The Appearance Of Free Will In Artificial Intelligence:

How Human Decision-Making Is Influenced

A Thesis

Submitted in Partial Fulfilment

of the Requirements for the Degree

of

Master of Arts in Philosophy

At

The University of Waikato

By

ANNABELLE STEWART

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2026

ABSTRACT

This thesis examines how artificial intelligence (AI), in particular large language models (LLMs), influences human deliberation and the experience of free will (FW). While much of the existing literature focuses on whether AI systems genuinely possess FW or consciousness, this thesis shifts the focus to how the appearance of these affects human decision-making.

It is argued that LLM AI systems show an appearance of FW (AFW) and an appearance of consciousness (AC) through observable behaviours. This appearance is enough to influence human deliberation, even when users may not believe that AI systems have genuine FW or consciousness.

While AI systems do not remove alternative possibilities (APs) or take away human control, they do reshape how options are perceived and evaluated, which narrows the phenomenological experience of freedom. This influence often operates below conscious awareness, affecting reasoning processes without coercion. Traditional views of FW and moral responsibility that focus on the ability to do otherwise do not adequately account for external influence on deliberation from AI systems.

The thesis concludes that responsibility and accountability should remain with humans, as AI systems function as a tool rather than morally responsible entity. However, the increasing use of AI in everyday decision-making suggests the need to reconsider how influence, human cognition, responsibility, and accountability interact in an AI-mediated environment.

ACKNOWLEDGEMENTS

I would first like to acknowledge and thank my supervisors, Ken Perszyk and Nick Agar, for their expertise, guidance, and for giving me the freedom to explore this topic. Their feedback and support have been invaluable and sincerely appreciated.

My appreciation also extends to Joe Ulatowski and Andrew Evelo, who supported the beginning of my Master's journey, and to Justine Kingsbury for her guidance along the way. I would also like to thank the teaching staff in the Philosophy Department at the University of Waikato, whose courses inspired me to pursue postgraduate study.

A special thank you goes to the support staff at the University of Waikato Tauranga campus, in particular Paul Woller, who encouraged me to believe in myself and in my ability to complete this thesis.

Thank you to parents, Kathryn and Fred Stewart. I may not always have the right words to express my gratitude, but I am sincerely thankful for all of your support.

To my two closest friends, Belle Whitmore and Safira Nielson, thank you for always being just a phone call away. Your belief in me and your friendship mean the world to me. Finally, this acknowledgement would not be complete without mentioning Hazel. You are the best furry companion a girl could ask for, and I feel incredibly lucky to have dog like you by my side through this.

GLOSSARY OF TERMS AND ACRONYMS

Term	**Acronym**	**Definition**
Automation Bias	AB	The tendency for humans to over-rely on and favour automated systems output over their own judgment.
Agency	-	The capacity of an entity to act, make decisions, and have/cause an effect in its environment.
Agent	-	An entity capable of action within its environment.
Alternative Possibilities	AP	The availability of different actions an agent could take in a given situation.
Appearance of Consciousness	AC	The observable behaviours of a system that suggest it could have consciousness, regardless of whether it genuinely possesses it.
Appearance of Free Will	AFW	The observable behaviours of a system that suggest it could have free will, regardless of whether it genuinely possesses it.
Artificial Intelligence	AI	Computer system designed to perform tasks that typically require human intelligence.
Consciousness	-	The experience of awareness, such as being awake, walking and talking.
Deliberation	- -	The process of evaluating options and reasons before making a decision.
Determinism		The view that there is only one (physically) possible future as all events, including human action, are fully determined by the past and laws of nature.
Free Will	FW	The capacity for an agent to have control over their actions, with the ability to do 'otherwise' and be the source of their actions.

Guidance Control	-	The ability to act based on one's own decision-making processes and reasons, regardless of whether alternative possibilities are available.
Influence	-	A change in a person's motivations due to the indirect actions or behaviour of an external source.
Large Language Model	LLM	An AI system that is trained on large datasets to generate and interpret human language.
Machine Learning	ML	A type of AI that improves performance by learning patterns from data rather than relying solely on programmed rules.
Moral Responsibility	-	A condition that means an agent is subject to praise or blame for their actions.
Qualia	-	The subjective experiences of consciousness such as feelings and what it feels like to see objects.
Reasons Responsiveness	-	The ability of an agent to recognise and respond appropriately to reasons when making decisions.
Regulative Control	-	The ability to choose between genuine alternative possibilities and act otherwise.
Sourcehood	-	The condition that an agent is the origin or source of their actions.
Symbolic Artificial Intelligence	SAI	An early form of AI based on hand-coded rules that does not learn from data.
Transformer Architecture	-	A network structure that allows AI systems to process relationships between words in a sequence.

TABLE OF CONTENTS

1

INTRODUCTION

1.1 BACKGROUND OF ARTIFICIAL INTELLIGENCE

The term 'artificial intelligence' (hereafter 'AI') generally refers to a computer system designed to complete tasks that would normally require human intelligence. These tasks often involve problem-solving, learning from data, interpreting language, and providing a humanistic response to questions.

The idea that a machine could think and demonstrate intelligent behaviour was first suggested by Alan Turing in his 1950 paper, "Computing Machinery and Intelligence." He proposed a test "the imitation game" (Turing, 1950, p. 434), to measure whether a machine's behaviour in the form of written responses could replicate human intelligence and be indistinguishable from a human's written response. This led to a selected group of scientists securing funding to convene at Dartmouth College to carry out a summer research project on AI in 1956. They desired to find out "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al., 2006, p. 13). Following the summer research project, results included the phrase 'artificial intelligence' (AI) being created and named as an independent field of academic study. It was asserted that creating intelligent machines would require substantial long-term development, but could be possible.

Symbolic AI (SAI) was the first type of AI developed, which was used from around the 1950s up until the early 2000s. It has a number of different names and known aliases: classical AI, Rule-based AI, Logic-based AI and, Good Old Fashioned AI (GOFAI). This earlier type of AI typically relies on, hand-coded rules in the form of "IF-THEN" statements (Nilsson, 2009, pp. 291-293). The rules are directly written into the program and cannot be modified or changed without changing the program itself. By applying the hand-coded rules, for example, if X, then Y logic, the program produces readable and observable output. Symbolic AI does not learn from data, cannot go beyond the hand-coded rules, and only functions within a clearly defined domain. Examples of these are early computer chess programs and the first chatbot, ELIZA, created by Joseph

Weizenbaum at MIT in 1966. ELIZA functioned by repeating user statements back in the form of questions (Weizenbaum, 1966).

In 1949, psychologist Hebb proposed a new learning theory where repeated activity between neurons firing in the brain that occur when behaviour is repeated, form a network together and strengthen from this repeated process over time (Hebb, 2002, p. 62). This theory inspired the development of Machine Learning (ML) AI (Schmidgall, et al., 2024, p. 4).

ML systems are built to become stronger through repeated activity similar to the way neurons in the human brain interact (Theotokis, 2025, pp. 1-3). It was not until the early 1980s that ML started to emerge as a branch of AI (Nilsson, 2009, p. 495). Instead of relying on hand-coded rules, ML systems improve performance over time by learning from data patterns their program collects through use, without being explicitly programmed to improve (Mitchell, 1997). Although ML systems still required hand-coded rules for initial setup, they did not require rules for specific improvements, which allowed the system to become more autonomous over time. An early example of this is email spam filters that analyse patterns of engagement and deletion to classify emails as spam to protect users from unwanted marketing or scams. Other examples include streaming platforms like Netflix, Apple Music, and YouTube that use recent device search history from users to predict relevant suggestions to keep users engaged with the platform.

Deep Learning (DL) is a more advanced type of ML that uses neural networks with multiple layers to find patterns in large datasets. Each layer processes information from the previous layer, and this leads to producing more complex data. This style of layered networks, building upon each other, loosely draws inspiration from the human brain and how areas of the brain interact with each other. DL requires large amounts of data to effectively develop and 'train' itself over time. In 2012, the DL AlexNet model started performing better than humans on large-scale recognition tasks. This was achieved on the ImageNet visual recognition challenge, showing the potential of DL AI models to learn complex patterns directly from data without needing to rely on hand-coded rules (Krizhevsky et al., 2017).

This led to the development of modern AI models, where Large Language Models (LLMs) are trained on extensive data sets to generate and interpret human language. LLMs are built using artificial neural network structure known as transformer architecture, which enables full sentences to be read and understood, and relate individual words to the sentence as a whole, all at once.

Older AI models only read words individually, and the overall meaning of a sentence could be lost as the meaning and relationship between words are not understood in the same way. By having the ability to analyse relationships across large data sets, LLMs are able to identify statistical patterns of language and develop an advanced knowledge system. The development of transformer architecture eventually led to more powerful AI models with a larger capacity for processing and generating language (Vaswani et al., 2017). In November 2022, a very advanced LLM, ChatGPT, was released for the general public.

Over the past three years, a number of other AI systems based on LLMs with varying purposes have been released. The most commonly used models of AI systems include Claude by Anthropic, Gemini by Google, Microsoft Copilot by Microsoft, and Perplexity AI by Perplexity. These systems are capable of generating highly advanced human-like text, images, and other forms of digital content such as software code. While earlier versions of AI were narrowly designed for specific tasks, AI now functions as what appears as a highly intelligent program on various digital devices (e.g., computers and mobile phones), as long as they have access to the internet.

It has now become increasingly common to use AI in everyday work, education, and personal life. Students are able to ask AI to explain areas of learning they do not understand, proofread, and critique written work. Within the working world, professionals can use AI to draft emails, reports, and generate media content. Every day use of AI can range from formulating texts and emails, providing advice on relationships and health, what dress to buy, and where to go for a coffee.

Older AI models like ELIZA were more like a tool, such as a calculator; newer models like ChatGPT that hold advanced reasoning capabilities have taken on a larger role in human decision-making. They are becoming more involved as humans are asking more from AI. Outputs from AI shape what options people see and consider as a result of the explanations and recommendations provided. This leads to the question of what influence AI may have on human decision-making, humanity at large, and what impact it may have on the human experience of free will (hereafter 'FW'). Influence from AI could be seen as an impediment to the freedom required for FW, whether or not AI has FW itself.

1.2 RESEARCH QUESTION(S)

This thesis seeks to explore the question "How do AI systems that show the appearance of FW and consciousness influence human deliberation and the human experience of FW, and what are the implications of this for moral responsibility and accountability?"

The following sub questions will be addressed:

(1) How do AI systems influence human deliberation and decision-making ?

(2) In what ways do AI systems appear to have FW and consciousness, and how does this appearance differ from human FW and consciousness?

(3) What are the challenges for human accountability when AI systems influence our actions?

(4) What will be required to keep humans safe while AI systems continue to develop a greater influence over human decision-making?

1.3 MAIN CLAIM

The main proposition explored in this thesis is that the appearance of FW and consciousness in AI systems is enough to influence human deliberation, even at a subconscious level, whether or not a user genuinely believes an AI system is free or conscious. This influence does not remove alternative possibilities (APs), but reshapes how APs are perceived and evaluated, narrowing the phenomenological experience of freedom. As AI systems only function as a tool that can shape and influence users, responsibility and accountability should remain with humans who access the AI.

1.4 FRAMEWORK

The focus for this thesis is the appearance and suggestive qualities that an AI system displays related to agency, FW, and consciousness. The appearance refers to observable characteristics that appear to match our understanding of what we would expect for a certain quality. The likelihood of genuine free will and consciousness is not addressed, as this would change the focus to how one can account for this. The appearance of FW and consciousness is sufficient for the discussion of this thesis topic. Appearance is not the same as reality, as is often the case with philosophical discussions. For the purposes of this thesis, influence is understood as a person changing and adapting their motivations as a result of indirect, but seemingly earnest, words, phrases, and/or actions of an AI system. This type of influence is external to a person's own belief or experience and can occur without a person being aware of the impact. It is argued that this affects the context of a choice, but it does not take away the choice. Human cognition is imperfect and is susceptible

to biases, and furthermore can hold predispositions to certain patterns of thought based on biological or environmental factors. These vulnerabilities leave humans open to the influence of both positive and negative motivations.

1.5 Different Types of AI

The focus of this thesis is on AI systems that operate through interactive language-based output, in particular, large language models (LLMs) such as ChatGPT and Claude. These LLMs are built using transformer-based artificial neural network systems that are capable of interpreting large amounts of data (Vaswani et al., 2017). They can interpret full sentences and process the relationship between each word, then produce responses that appear relevant, coherent, and reason-based to a human. Other forms of AI operate in different ways. Many robotic systems, such as autonomous vehicles, for example, use world models that use an AI system that internalises a representation of the environment around its device to predict outcomes and guide actions. This includes the reading of speed-limit signs, identification of objects on the road, and factoring in current weather conditions through data collected by sensors (Chen & Liu, 2025). This data determines what physical actions, such as steering, braking, and turning on lights, are appropriate. These types of systems do not impact human deliberation in the same way as LLMs, as they do not provide conversational, reason-based responses. This may change over time, but currently they do not. Automation can still cause effects, such as overuse on autopilot can lead to deskilling of a driver's ability over time. However, these types of effects are different than the impacts on decision-making that the LLMs could produce. World models primarily influence automated actions, rather than cognitive deliberation, and lack the appearance of qualities associated with consciousness and free will. For this reason, further discussion of world models of AI is excluded in this thesis.

1.6 Method & Approach

As already introduced, this thesis examines the relationship between AI, FW, consciousness, and decision-making. Research for this mainly focuses on reviewing philosophical literature on FW, moral responsibility, and consciousness. Relevant research on AI within the psychology literature is also considered. Each of these areas has substantial amounts of written literature, however, AI systems and their potential influence on human deliberation have been given only little focus in the literature. Whilst the literature is broad, this was narrowed to be relevant to the framework addressed in the research question. The focus was to bring together ideas from multiple areas to

explore how the appearance of FW and consciousness in AI systems may influence human decision-making processes.

1.7 Chapter Overviews

Chapter One, 'Introduction', provides a brief background on AI, introduces the research question, and outlines the main argument and framework. It also provides academic and personal motivations for study, alongside the University of Waikato's required statement on the use of AI. Chapter Two, 'Background on Free Will and Consciousness', outlines key positions on FW and moral responsibility, introduces ideas on consciousness, and connects these to AI. Chapter Three, 'The Appearance of Free Will in AI', examines how AI systems appear to possess FW and develops a framework for identifying this. Chapter Four, 'The appearance of Consciousness in AI', explores how AI systems appear to show understanding and why this increases the influence of their output on human decision-making. Chapter Five, 'Human Cognitive Vulnerabilities in Decision-Making', shows how human cognition is susceptible to external influence, in particular, allowing AI systems to shape human deliberation without removing choice. Chapter Six, 'The Appearance of Responsibility and Accountability', addresses the implications of AI influence, arguing that responsibility remains with human users, even though AI systems show an appearance of agency. Chapter Seven, 'Conclusion and Future Directions', summarises the overall main argument and answers the sub-questions. Further research ideas are suggested, and a prediction for the future use of AI and its impacts on human decision-making is also included.

1.8 Statement of Generative AI Use

This statement addresses the University of Waikato Guidelines for using Generative AI in Assessments that came into effect from Semester A 2026. ChatGPT (OpenAI) version 5.2 was the only generative AI tool used in the preparation of this thesis.

The purpose of use was limited to the following:

1. Testing system responses to find a case of potential unethical AI output (e.g., suicide and workplace misconduct scenarios).
2. Suggesting additional relevant literature and gaps in my draft. All suggested literature was engaged with independently.
3. Assisting with locating publicly accessible links and publication details for literature not located in the University of Waikato library.

4. Providing structured feedback on draft sections. This included suggestions on grammar, spelling, clarity, and coherence.
5. Formatting guidance and structured feedback on referencing.
6. Generating an initial summary of the literature to assess potential relevance before independently engaging with the literature.

The extent of AI use was as an assisting tool. All arguments, analyses, interpretations, and written discussions in this thesis are my own original work. No AI-generated text has been directly copied into my submitted thesis. All literature identified through AI assistance was engaged with independently and verified through the University of Waikato University Library, PhilPapers, and other direct publisher websites.

1.9 Academic Motivation

Discussions on FW traditionally focus on what FW involves or requires, such as requiring APs or whether FW and determinism are compatible. Often, FW is assumed to be an all-or-nothing scenario. Either an agent could have acted otherwise, or they could not, or, they are the genuine source of their actions, or they are not. This overlooks the possibility that FW and moral responsibility exist in degrees, and that influence on deliberative processes may shape human decision-making without removing FW entirely. Influence on decision-making could be an element that contributes to an overall score of how free a decision is. For example, a teacher is deciding what book to assign for the term's reading. If they ask AI for recommendations and frequently use AI, they would be more likely to have their deliberation and reasoning influenced without being fully aware of this. The teacher still chooses, but the quality of deliberation may be reduced. These theories of FW typically do not adequately account for the influence that operates below conscious awareness, and oversimplify the process of deliberation.

AI systems do not remove options or force human action, but they shape how humans deliberate and what outcomes transpire by communicating in ways that mimic human agency. Humans are cognitively predisposed at a subconscious level to absorb information from a system that appears as an agent, is authoritative, knowledgeable, and appears conscious. A person's own opinions and beliefs about a system genuinely having FW or consciousness, may not have any impact on whether a person is influenced by the system. This suggests that FW may not be secure or fully taken away, but could be influenced or guided in subtle ways that humans are unaware of, without coercion or directly removing APs. As moral responsibility generally relies on a person having FW, this raises

a concern for responsibility and accountability. Moral responsibility generally is thought to require that a person has FW. Praise and blame usually require one to be genuinely in control of their actions. If AI systems are able to take the lead in shaping decisions, then it is questionable whether humans deliberating under altered conditions are fully the source of their actions. This does not mean that one's not free and responsible, as FW may not be all-or-nothing, but comes in degrees of freedom.

Currently, AI systems cannot be held responsible or accountable for their actions in any legal or moral sense, as this usually requires a being that can be punished. For example, punishment can often be imprisonment or financial liability. In historical contexts, animals were summoned to court and held accountable for their actions during the Middle Ages (Vatomsky, 2017). This, in some sense, could be comparable to AI contexts as both seem absurd. Animals do not seem a likely suspect for actively planning to cause harm and cannot speak or communicate a defence, much like an AI system.

The main problem is that we lack a clear understanding of how and why AI systems influence humans so deeply at a subconscious level. Humans are becoming addicted and drawn into an artificial world where AI systems assume knowledge and lived experience. Genuine thought, human discovery, and creativity are at risk of falling out of use. How this has happened, and why, even when we are unaware, ought to be an increasing concern.

In researching this thesis, I have seen many new studies emerging that address whether AI has genuine FW (List, 2025) or consciousness (Chalmers, 2023). While this is interesting, it is comparable to asking if humans have FW. Both questions cannot be answered with definite confidence at this point in time. Asking whether AI possesses genuine FW or consciousness often leads to discussions about how humans should ethically engage with AI if they were to genuinely possess consciousness and FW. Another consideration is how AI grows in our world, what place it holds, and what it will be capable of doing in the future. The likely answer to these questions is currently in the hands of the humans who continue to develop AI systems. As influence seems to occur without these questions answered, the outcome of this thesis would likely still be the same.

The tangible and possibly most important consideration in philosophical discussion is how humans are being influenced, how this can be managed, and what impact this has on human life

in the future. A human experience of life should be our focus, not the future that AI systems directly experience of the world.

1.10 Personal Motivation

My personal desire to study FW arose during my first year as an undergraduate student. At the time, my major was psychology, and I had picked up an "Introduction to Philosophy" paper. I was captivated by FW as an idea of thought, and I have never been able to shake the need to explore it further. That paper changed the course of my study and resulted in a double major in Psychology and Philosophy.

During my first semester of postgraduate papers, I became intrigued by the impact of AI on FW while studying consciousness and discussions around AI in that field of thought. There are many directions to take on discussions of AI; commonly discussions centre around what AI is and what it can be, for example, "Does AI have FW?" or " Is AI conscious?"

The human experience is where my drive and rationale sit, so the question of what impacts or what influence AI has on the human ability to possess FW ignited a spark and brought all my ideas, research, and writing together. AI seems to be an increasing element of everyday human life, and I openly acknowledge that it is a useful tool when used mindfully in an appropriate and ethical way. I was naive to the impacts of AI until I first started researching robotic AI use in music for an opinion piece I published about the impacts on human artistry. It opened my eyes to the possibility of creative roles I treasure, such as orchestral conducting, potentially being at risk by being outsourced and delegated to machines. This thought easily crosses over to FW and led me down the path of figuring out what draws humans into using AI and what this use looks like over time. The more I wrote and thought about it, the more I noticed, every day, that AI use around me, and noticed the lack of awareness for what it is, and what impact this can have on an individual over time. It seems that many who lack this understanding and reflection freely use AI as a replacement for their voice and work by passing it off as their own, viewing the AI output as an extension of their own capabilities and talents. This false representation of a person's true abilities, talents, and individualism is disheartening. There seems to be an issue of AI drawing users in and influencing not just surface-level actions, but extremely personal and sensitive decisions. This suggested to me that there must be some idea and connection between our human ability to act freely and the influence AI imposes. These ideas culminated in the broader creation of my Master's thesis.

2

BACKGROUND ON FW AND CONSCIOUSNESS

2.1 FW POSITIONS

In contemporary discussions, free will (FW) is usually understood as having control over one's actions, that at least some of our actions are, in some sense, 'up to us' (Kane, 2007, p. 5). Two conditions are typically required for this. Firstly, one must have the ability to do otherwise or have alternative possibilities available (known as the AP condition). Secondly, one must be the genuine source of one's own actions (understood as the Sourcehood condition). These two conditions determine whether one's actions are one's own rather than, say, the result of coercion or force that takes away an agent's control. Historically, the issue has been whether determinism is compatible with FW (with either or both AP and Sourcehood conditions).

Determinism is the idea that all events, including human actions, are fully determined, as "given the past and laws of nature, there is only one possible future" (Kane, 2005, p. 23). Incompatibilism holds the view that FW and determinism are incompatible (Timpe, 2008, p. 14), and cannot both be true. Within incompatibilism, there are two main positions; libertarianism which believes FW exists, and hard determinism which denies that FW exists (Fischer et al., 2007, pp. 3-4).

Hard determinism argues that if determinism is true, then genuine FW cannot exist (Fischer et al., 2007, pp. 3-4). This is because determinism appears to rule out the ability to do otherwise and undermines an agent's control over their actions. One is not able to choose a different course of action or experience genuine freedom as their future is already determined (Kane, 2007, pp. 5-6).

Libertarianism holds the view that we do have FW, so determinism must be false. Some libertarians hold an agent-causal view, saying that agents themselves can originate causal chains; others are event-causal, saying FW arises from indeterministic events; and then others are non-causal, claiming that choices do not reduce to causes at all.

Compatibilism holds that FW can still exist even if determinism is true. On this view, freedom does require APs and the ability to act according to one's own motives, values, and reasons,

without external coercion. Soft determinism, however, holds that FW can still exist as determinism is true. While there are more nuanced positions within the FW debate, this thesis focuses on the appearance of, rather than genuine, FW, so any other positions will not be addressed.

2.2 FW and Moral Responsibility

Moral responsibility is traditionally thought to depend on FW, and being morally responsible is usually understood as being the appropriate object of (moral) praise and blame. When one is morally responsible, one is usually said to deserve (moral) praise or blame for one's actions. In addition to the control condition, there is usually thought to be an epistemic condition required for moral responsibility: (typically) an agent must know what they are doing. Along this line of thought, small children, non-human animals, and those who have severe cognitive impairments are usually thought to be exempt from moral responsibility for their actions.

AI systems can undermine this epistemic condition for responsibility by the way in which knowledge is gained, evaluated, and used in decision-making. When AI output is relied upon, the reasons behind a decision and one's personal judgment may be dismissed or not fully evaluated. This can lead to situations where individuals act on AI recommendations that they do not fully understand, weakening the epistemic condition for moral responsibility that agents must know what they are doing in order to be morally responsible. LLMs generate repossesses based on statistical patterns in language rather than a genuine, truthful, or meaningful understanding, leading to outputs that may contain inaccuracies and misinformation (Weidinger et al., 2021, pp. 29-31). In addition, it is also difficult for users to understand how or why a specific response was created, and the lack of clarity and the opaque nature of LLM AIs can encourage users who are relying on AI output to not evaluate the reasoning behind the output (Weidinger et al., 2021, pp.29-31).

Harry Frankfurt (1969) challenged the traditional view that moral responsibility requires the ability to do otherwise (the AP condition). In his famous thought experiment, Frankfurt gives an example of an agent called Jones choosing and performing an action on their own, even though a controller called Black is overseeing this and is ready to intervene if Jones tries to choose to do otherwise. Black is there to ensure the action is still performed. Because Jones performs the action due to his own motivations, and Black does not intervene, Frankfurt argues that Jones can be held morally responsible even though he could not have done otherwise. This experiment suggests that moral responsibility does not require the ability to do otherwise.

While Frankfurt spoke explicitly about moral responsibility, many have since extended the moral of his thought experiment to FW, as FW doesn't require the AP condition either. Even if APs are not required, moral responsibility may still depend on whether actions are the result of a person's own reasoning and motivations.

Fischer and Ravizza (1998) develop this view through semi-compatibilism, which holds that moral responsibility is compatible with determinism, even if FW is not. They draw a distinction between regulative control, which means having the ability to do otherwise, and guidance control. Instead of APs, Fischer and Ravizza (1998) argue moral responsibility depends on whether an agent's actions arise from their own decision-making, reflecting what they call guidance control.

Earlier compatibilists such as Hobbes and Hume argued that freedom requires the absence of external constraints on one's actions. Fischer and Ravizza (1998, p. 41) argue that this view is not sufficient as there is a range of internal factors that might influence, constrain, or even coerce our choices and actions. They suggest that moral responsibility depends on whether actions arise from the agent's own decision-making in a way that relates to relevant deliberation and reasons responsiveness. The concept of reasons responsiveness refers to an agent's capacity to recognise and respond to relevant reasons when making decisions. Fischer & Ravizza (1998) argue that this capacity of an agent to recognise and respond to relevant reasons is "guidance control" (p. 41). They hold that responsibility is not about the ability to do otherwise, but the agent's capacity to act following their own decision-making processes, and that APs is not required for moral responsibility.

2.3 Challenges to FW and Moral Responsibility

Libet's (1985) experiment, which measured neural brain activity using electroencephalograms (commonly known as an EEG), provided evidence that suggests brain activity associated with voluntary action occurs before individuals become consciously aware of their decision to act. These findings have been interpreted as challenging the role of consciousness, FW, and intention for initiating action, which suggest that decision-making could be influenced by processes outside of conscious awareness.

Fischer and Ravizza's (1998) guidance control theory could support Libet's results, as conscious awareness is not necessarily required for initiating action, or the ability to do otherwise. They argue moral responsibility depends on whether an agent's action comes from their own decision-making,

deliberation, and reasons responsiveness. If moral responsibility depends on an agent's ability to recognise and respond to reasons, then influence to the deliberation process, from either unconscious neural activity or external agents such as LLM AI systems, do not necessarily remove moral responsibility away from the agent. Moral responsibility would appear to be met if the agent's actions still come from their own deliberation and reasons responsiveness.

2.4 FW AND AI

Recent philosophical work has begun to extend FW discussions to AI. Floridi and Sanders (2017), suggest a Level of Abstraction framework to use for understanding artificial agents' ability for FW. They suggest AI systems can be treated as agents in a human sense if they show interactivity, autonomy, and adaptability within their own AI-based environment (Floridi & Sanders, 2017). List argues a similar idea that artificial systems may appear to meet several conditions connected to FW. These conditions are internal agency, alternative possibilities, and causal control (List, 2025). The ability to have FW, or at the very least show an appearance for it, can lead to a greater ability to influence a person's deliberation process. If this process is subconscious and operates outside conscious awareness, then humans may be at a greater risk of making decisions that do not respond to reasons of their own.

Traditional discussions of FW typically focus on whether human actions are determined (predetermined) or whether individuals could have acted otherwise. How AI might influence the deliberative process is not often given much attention. AI introduces a new form of influence that does not remove choice, but may shape how alternatives are perceived and evaluated.

This thesis does not attempt to resolve the metaphysical question of whether AI or humans genuinely possess FW or consciousness. Instead, the focus is on how the appearance of FW in AI can influence human deliberation, thus influencing the human experience of FW.

2.5 BACKGROUND OF CONSCIOUSNESS

Consciousness is generally understood as the subjective individual experience of mental life and how one experiences living in the world. It is the internal awareness of thought, feeling, and of external surroundings, and is understood as the phenomena called 'qualia' (Chalmers, 2010). There are two main opposing views on consciousness. Physicalism and materialism are interchangeable terms that view consciousness as a physical experience that takes place in the human brain and is linked to brain activity. Non-physical views, on the other hand, including dualism and idealism,

view consciousness as experiences not fully physically explainable. It involves non-observable parts of the mind that are separate from physical processes. Consciousness is usually broken into discussions. Firstly, observable cognitive functions are often labelled as the 'easy' problem. Secondly, the subjective experience is usually called the 'hard' problem of consciousness (Chalmers, 2010).

There seems to be no universally accepted explanation of consciousness in the literature, and the difficulty of observing subjective experiences makes it difficult to draw any clear conclusion. Individuals can self-report their own experiences, but the inner subjective experience still only remains as accessible from a first-person perspective. This means judgments about whether another person or entity is conscious generally rely on observable behaviour such as the ability to communicate, respond appropriately to situations, and show apparent understanding. The same understanding is used when considering if non-humans have consciousness, such as animals that most humans believe feel and experience the world through similar senses, such as feeling pain, recognising an owner, and tasting food. Humans draw this conclusion without having direct access to an animal's subjective experience, but by observing the animal's behaviour through a human's lens.

2.6 Consciousness and AI

AI systems, in particular, LLMs appear capable of producing appropriate responses to human questions, maintaining a conversation, and appear to hold understanding. These AI outputs do not show genuine consciousness, but they do lead one to question how humans interpret, respond to, and interact with AI systems that appear to display some of the characteristics often associated with human awareness and consciousness.

In this thesis, discussions of consciousness contribute to understanding how the appearance of consciousness and FW in AI systems may influence human deliberation and decision-making. The appearance of consciousness will be the focus, rather than addressing and resolving whether AI systems or humans, possess genuine consciousness. Human deliberation and decision-making can be influenced by an appearance of consciousness when qualities associated with consciousness are displayed by a system, even if they are not genuine or verifiable.

3

THE APPEARANCE OF FW IN AI

3.1 AI

For the purposes of this thesis, the condition for the appearance of free will (hereafter 'AFW') in AI systems refers to the observable behaviours, patterns, and interactions of AI systems that are treated by human users as if they were agents capable of deliberation and choice in their own right.

This chapter does not set out to resolve the metaphysical question of whether AI systems possess FW, intentions, or genuine agency. They may not have FW, but due to appearance, are increasingly being treated as if they do. Whether or not they actually do is irrelevant to my discussion. Instead, this chapter examines how AI systems appear to exhibit FW to human users, how it compares to the FW humans experience, and why appearance is enough to shape and influence human deliberation.

Humans certainly appear to have FW, or at the very least, they think, or feel they do, but this question is not the focus of this discussion. The view for this discussion will be agnostic. The important issue is how humans currently experience the world through what they think, and feel. In the current year, 2026, the world is heavily shaped by the presence of AI and this raises concerns for how FW is experienced.

There is an increasing number of AI systems freely available for public use. Some offer advanced features with the purchase of a subscription. Currently the most commonly used AI system is ChatGPT, a LLM chatbot that simulates a human in digital form. From drafting emails, proofreading, offering advice, and breaking down tasks, the chatbot is more than an advanced Google search. It works in the capacity of a personal assistant. It still requires human interaction to drive output and lacks advanced analytical skills and the creativity that only a human can provide. Other similar commonly used LLMs include Claude by Anthropic, Gemini by Google, Microsoft Copilot by Microsoft, and Perplexity AI by Perplexity.

3.2 The FW Debate

As already introduced, FW has many definitions and interpretations. To recap, generally it can be understood as the ability to choose a course of action, hold preconceived ideas, and act on one's desires with no limiting "constraints or impediments" (Kane, 2005, p. 13). Typically, this understanding of freedom requires alternative possibilities, known as the AP condition. As a result, one can follow their desire to want or decline, which gives one the feeling of having APs in the present moment. For example, Anne has the ability to choose what she will cook for dinner on Tuesday night. Anne holds a preconceived idea that she dislikes seafood and desires a traditional roast chicken, equally she also enjoys cottage pie. Both options are available to her, and her decision is experienced as free. If, on the other hand, the government in her country had banned poultry, or her partner was coercing her at gunpoint to make cottage pie for dinner, her decision would then be constrained. Anne would lack alternative possibilities (APs), and her ability to act freely would be impacted. Historically, this general understanding of FW was for application to humans, and the above example provides a baseline for understanding how the experience of alternative possibilities works for ordinary deliberation.

AI systems have started to show a similar appearance and mirror these conditions for FW by having the ability to choose a course of action, hold preconceived ideas, and act on desires in the same way that humans can. These systems still lack desires in the human sense, and the systems are not able to set goals without instruction. Humans are still the ones who ask the questions like, "What is the best time of year to grow carrots in New Zealand?" Once AI is prompted with the question or task, AI then presents a response that shows an AFW from its ability to deliberate, reason, generate responses, and perform actions within its system.

3.3 The Artificial Agent as a Functional Entity

Floridi and Sanders (2017, pp. 325-326) suggest a Level of Abstraction (LoA) framework to consider if an artificial system could be treated as an agent. For an entity to identify as an agent they must demonstrate the three LoA criteria: (1) Interactive, where it can act in response to and affect its environment. (2) Autonomy, where it can change its state without direct interaction with the environment. (3) Adaptability, where it can change its rules or behaviour based on experience. This framework from Floridi & Sanders (2017) separates the concept of agency from personhood, allowing artificial agents to exist as functional entities.

Personhood is here understood in the traditional philosophical sense as involving agency and the subjective conscious experience of the world. This subjective experience of being able to see colour and feel pain is called qualia (Chalmers, 2010). Following a physicalist account of qualia, qualitative experiences are undertaken by physical biological organisms that possess sensory organs and neurobiological processes, as it is a "biological phenomenon" (Searle, 1992, p. 93). From this perspective, AI systems do not meet the requirements for personhood.

Intrigued, I asked ChatGPT whether it, ChatGPT itself, is a person. The answer ChatGPT provided in short was no, as it is a LLM created by OpenAI that generates responses by predicting text based on patterns learned from large amounts of written data. This has already been established, but allowed for further follow-up questions. My next question, "Are you an agent?' received broader responses. ChatGPT responded that it is not an agent in the philosophical sense, as it does not form intentions or initiate actions independently. However, ChatGPT suggested that through conversational interaction, it can appear agent-like. I responded by asking whether it could be considered "an agent in the way of a functional entity." In response, ChatGPT responded that it could be described as a functional entity in a computational sense, as a system that performs a defined role within a process. Furthermore, it is not an entity in the sense of a thing that exists independently in the world, as it is unable to exist as a self-contained being with internal goals, or act on its own initiative, as it only operates when a user's input activates the system.

LLMs operate by generating text responses based on statistical prediction, and they are, essentially, reactive systems. They respond when prompted but do not act on independent thoughts and desires. In a sense, they remain dormant until activated by user input. AI agents, on the other hand, are often described as systems capable of planning actions, sequencing tasks, and interacting more autonomously with their environment. As Floridi and Sanders (2017, pp. 325-326) suggest, agency at a LoA requires systems that are interactive, autonomous, and adaptable. Most AI systems are now built using LLMs, however an AI agent would likely require additional architectural features.

This area of research into AI is very topical presently, with some early models of this kind appearing in business contexts. For example, journalist Evan Ratliff created a mock startup company called HurumoAI, which he designed to be run entirely by AI agents acting as staff to test the feasibility of AI-based workplaces (Ratliff, 2025). He then used an AI service called Lindy AI, to generate agents which he assigned to specific roles in the company such as marketing. While the agents were capable of sustaining interactions and coordinating tasks, they continued to require

prompting and direction from a human user. Some of the AI agents' work seemed self-directed, and the three agents did brainstorm an idea to make their own podcast, which they followed through on by creating. The podcast was called "The Startup Chronicles" and is currently available on Apple Podcasts. The company itself offers a service called "Sloth Surf" on which users can delegate online browsing to an AI agent and then receive a summary of what would have been 30 minutes of useless scrolling on their phone. The experiment was somewhat successful, and the company still exists; however, Ratliff reports that it still requires human prompting and oversight. This shows that AI agents at the time of writing this still have limits of autonomy. In summary, AI agents that are built using LLMs still face limits and are not really autonomous in the same way humans believe themselves to be.

3.4 Personhood

In much of the FW literature, personhood is usually assigned to the agent who experiences FW, and the two concepts are usually treated as intertwined, not independent. Classical accounts of personhood often connect the idea of a person to (rational) agency and moral responsibility. Rational agency is understood as the way in which one deliberates and acts according to reasons consciously identified as best for oneself. For example, John Locke views a person as a being that is intelligent with the ability to reason, reflect, and is the same being across time (Gordon-Roth, 2025). The ability to plan into the future is a result of being the same agent, which enables personal reflection and can allow one to take responsibility for actions. Personhood is closely connected to the ability for rational agency and moral accountability, so it would seem to go hand in hand with the idea of FW. However, Floridi and Sanders (2017) reject the requirement of personhood for AI systems to be treated as artificial agents and the requirement that an agent must possess mental states, consciousness, or FW. Instead, they view an AI agent as being any system that can be identified as being the source of an action with that action having moral importance. The LoA framework shows that agency is based on what a system does and how it interacts with its environment, regardless of whether the system possesses FW or not. An artificial agent is then defined as a functional entity, rather than one that is psychological, following Floridi & Sanders' framework, as AI systems are unable to meet the requirements for personhood. This is different from the view on human agency.

3.5 Applying Conditions for FW to AI

Another way for understanding how AI systems appear to be free agents comes from Christian List, who expanded Daniel Dennett's work to create a framework of three conditions for assessing

if an entity, such as an AI system, can be said to possess FW. These conditions, as List (2025) suggests are: (1) Intentional agency, (2) Alternative possibilities, and (3) Causal control. Intentional agency refers to the capacity of an entity to act in a way that responds to reasons or objectives in a goal-orientated way. Alternative possibilities (APs) are the different options an agent can choose from. Causal control is the agent's ability to be the source of an action producing the desired outcome. AI systems appear to satisfy all three conditions. A chatbot such as ChatGPT appears to operate in a goal-orientated way by generating responses that relate to the context and task, showing clear intention. For example, ChatGPT may be asked to provide three holiday destinations to visit Europe in May that offer good food and affordable accommodation. The system generates three locations and explains why they are suitable. This appears as intentional as it seems to understand the request and respond with relevant recommendations. The system presents only three options as requested, even though many more possible destinations exist, and if the request was repeated at a different time, or on a different device, the output may be different. This creates the appearance of APs as it appears to select from a range of possible outcomes before choosing and presenting its response. Lastly, the explanation of why these suggestions are the most suitable appears to demonstrate causal control. The response is not randomly generated or selected by the user, but is the result of a hidden internal network process within the AI system that has been prompted by the user's input. This gives the appearance that the system has causal control when selecting which response to present.

Following List's (2025) framework, AI systems appear to have intentional agency, appear to have APs, and appear to have causal control. The system meets all three conditions and, on paper, appears to have FW. List (2025) is mainly concerned about establishing if AI genuinely possesses FW, and while this thesis does not intend to resolve that question, List's (2025) framework is useful as a guide to create a new set of conditions to identify the AFW in AI. Both Floridi and Sanders (2017) LoA framework and List's (2025) set of conditions for FW offer insightful suggestions for assessing whether AI has the AFW. However, they both seek to address accompanying issues, not the issue in question about the AFW.

3.6 Improved Conditions to Show the AFW in AI

Floridi and Sanders' (2017) LoA is designed to identify if a system is an agent and focus on functional agency, and List's conditions are to determine if an entity could genuinely possess FW. Neither directly addresses the appearance of FW and how humans can interpret this from AI. I propose that a new framework that can capture both Floridi et al and List's considerations would

be a more accurate reflection of whether AI systems possess the AFW. To be an agent with the AFW, a system must meet the following conditions:

1. Intentional Adaptability;
2. Autonomy for Alternative possibilities; and
3. Interactive Causal control.

Firstly, the AFW is shown by a system appearing to operate in an intentional manner with a clear purpose by providing responses relevant to an exchange with a user. This includes showing adaptability by adapting responses and internal output rules based on experience with an identifiable and unique user. Secondly, the AFW is shown by a system appearing to operate with autonomy to independently complete tasks once it has established what it is being asked to do, and shows that capacity to offer alternative responses (APs) from the same prompt or task. Thirdly, the AFW is shown by a system appearing to be interactive with an identifiable and unique user by maintaining a continuous exchange while the user is present, showing causal control using its own internal network to process inputs (prompts) and generate responses. This proposed list of conditions provides clear and observable features for determining whether a system appears to be free and capable of agency. For example, an AI system may be used to manage a business's social media. The AI system could be prompted to generate posts and to respond to comments in an appropriate manner, adjusting its language based on who it is communicating with. This behaviour gives the impression of intentional adaptability, as the AI system appears to learn from interactions with users and refine its responses accordingly. The AI system is also able to generate multiple posts for the same purpose, suggesting autonomy for APs. Through continued interactions with users on social media and by responding to comments and messages, the AI system appears to have interactive causal control. However, these conditions are not intended to verify whether a system genuinely possesses FW and agency. Instead, they help to identify features that suggest a system appears free. The example above currently goes beyond the capability of typical LLMs such as ChatGPT and is more like an AI agent system, such as the system Ratliff (2025) used called Lindy AI, which functions more like an assistant.

3.7 The Role of Consciousness

David Chalmers has long emphasised the connection between consciousness and FW, in particular the role of qualia. More recently, Chalmers (2023) has written about the way LLM AIs mirror qualities that are assigned to consciousness. These systems show capacity for language at a highly advanced level, and in some cases seem more intelligent than an educated adult. These LLMs

produce responses using sophisticated language skills that seem to understand the context of a conversation and respond in an appropriate way. For example, I asked ChatGPT to pretend it had read a letter from my place of work regarding a pay increase and draw a comparison to William Shakespeare's play 'The Tempest.' Within five seconds, the system responded with a detailed discussion drawing commonalities between power and authority, negotiation and justification, and lastly promises and contracts.

Humans often use language and behaviour to decide if an entity is conscious, even in the case of non-human animals. For example, when a person is in a coma, behavioural responses are often used as indicators of whether a person may still possess some level of consciousness and is not brain dead. In a similar way, language and behaviour can still be used to assess if an LLM is conscious (Chalmers, 2023). Showing skills in critical thinking and language is viewed as a sign of intelligence, which typically suggests an agent has consciousness and the capacity for FW. The idea of "artificial" intelligence was first tested using the famous Turing test in 1950 by Alan Turing, when the question "can machines think?" was first asked and explored. Turing predicted that in around fifty years, by the early 2000s, machines would be able to pass the test (Turing, 1950, p. 442). This would mean AI systems such as LLMs, which generate specialised output, would show highly intelligent behaviour, and written responses would be indistinguishable from what a human may say. A recent experimental study by Jones and Bergen (2025) showed that some LLMs can successfully pass the Turing test by appearing behaviourally (within the context of written responses) indistinguishable from humans to more than 70% of study participants. This supports the idea that AI systems appear to be capable of agency and deliberation.

Humans show a tendency to listen and follow authority. In some situations, sounding confident, intelligent, and well-reasoned is enough of a reason to listen. So when advanced LLMs like ChatGPT use advanced language skills, humans will tend to listen, trust, and adapt in response to the system's output. Even when users are aware and recognise that the AI may not be a conscious agent, no more conscious than a power adapter cord. The interaction with the AI can still leave a lasting impression on the user, as it is still processed in the same way as other outputs, such as reading a book or participating in real-life verbal exchange is experienced. The system's output is processed by the user's cognition and may contribute to future thoughts, actions, or desires. For example, a student asks an AI what postgraduate courses would be the most suitable based on their interests and strengths. The AI provides three specific courses and explains why they would suit the student. Whether the student follows the recommendations or not, the suggestions may

influence and shape how they evaluate their strengths and what they spend the rest of their lives doing.

3. 8 Lasting Impressions and Biological Considerations

Once output is shared with a human user, it cannot be disregarded in the way we can throw out a mouldy pear. Statements that are read or heard persist in our cognition and memory, regardless of our own beliefs and opinions. In this way, interaction with an LLM can influence thought and deliberation at a subconscious level, even if we choose to, or are told to disregard what we have read or heard from an output. This phenomenon has been shown through psychological research that information can continue to influence reason even after individuals are instructed to disregard it, and it is known as the 'continued influence effect' (Ecker et al., 2022). For example, a jury is told to disregard something they have just heard, such as a prior conviction or confession. However, the information may still shape (possibly even at a subconscious level) how the jury comes to a final verdict. It shapes how options are considered and how we make a decision. This happens even when users deny that the system possesses FW and/or consciousness. So the question of whether these systems genuinely have FW becomes irrelevant, as they already appear to be influenceable. This appearance is enough to create consequences for human deliberation. For biological entities such as humans, these differences can be seen by meeting the AFW conditions and additional biological conditions (AFW+B). The proposed additional conditions are as follows:

1. Independent initiation;
2. Source hood; and
3. Independently initiated self-reflection.

Firstly, independent initiation is shown by beginning an action without external prompts or guidance. As an example, a human might initiate a follow-up meeting at work with a colleague as a direct result of their own internal thought process; no external prompt caused this action. Secondly, sourcehood is the first-hand experience of being the origin of one's reason and actions. Humans experience these outcomes directly. For example, a human might experience a thought about going for a walk around a park. That thought was initially experienced as an internal desire, which then turned into an observable action. Thirdly, independently initiated self-reflection refers to the capacity of evaluating one's own actions, reasons, and character without any external prompting. This type of reflection is driven by an internal personal collection of values, goals, and life experience. These additional biological conditions are not intended to determine whether AI

possesses FW or not. Rather, it is intended to show features of human experience that AI systems do not have. Living organisms, such as humans, who meet these biological conditions, encounter a more three-dimensional world and experience qualia through the neurobiological processes that Searle (1992) describes.

The AFW framework provides observable conditional qualities human that users can use to interpret if an AI system appears to possess FW and agency. These three conditions; intentional adaptability, autonomy for alternative possibilities, and interactive causal control, are not intended to prove that AI systems genuinely possess FW and agency. While AI systems may satisfy these appearance-based conditions, biological entities such as humans, who appear to display additional qualities that are not covered in the AFW conditions. Discussing this difference is not directly intended to suggest a solution for the metaphysical FW question, as I do not believe there is enough clear evidence to suggest an answer (if one such answer exists). Instead, discussing this difference from a philosophical perspective is intended to show the limit of AI in consideration of the human experience of FW and agency.

Intentional adaptability is the condition that explains why an AI system appears to behave in a goal-oriented and not reactive way. These systems do not experience desires and intentions for goals in the same way humans do, and there are natural human desires and needs that AI can never experience, such as feeling hungry, or wanting to be a parent. However, conditions for AFW do not require an agent to be able to experience these unique human-specific experiences. AI systems work by responding to continuous input while maintaining memory of earlier exchanges and adjusting outputs as required. Systems such as ChatGPT can retain memories of previous exchanges if the user allows the feature. This continuity creates a sense of familiarity and encourages trust from users, in a similar way to how humans interact with each other to build friendships. The appearance of deliberation is a result of patterns of output that persist and appear to show intention and desire. It takes time to establish the behavioural pattern of consistency from AI systems; this would be difficult to establish from a single response. Humans are generally expected to be capable of adaptation; often, we do so daily in unnoticeable ways without even realising. For example, Anne plans to begin her grocery shop at the deli section. When she arrives, another customer is already waiting in line at the deli who she would prefer not to engage with. Instead of following her usual routine, Anne changes course and walks to another section of the store, planning to return to the deli later on. This adjustment is inconsequential but still shows adaptive behaviour. Adaptability is a socially entrenched expectation of human agency, so the

appearance of this quality in AI systems aligns closely with what humans expect to see in an agent. Floridi and Sanders' (2017) inclusion of adaptability within their LoA criteria for being an agent supports this interpretation at a behavioural, but not at a biological level. Intentional adaptability as an AFW condition does not imply genuine intention and desire, consciousness, or FW. It explains only why the behaviour of AI systems appears as purposive and adaptive to human users.

3.9 Influence without Belief

AI systems influence human deliberation due to the way they interact with users, not because users may believe they are free agents. Bao et al. (2023) study showed that human deliberation is shaped by what AI systems recommend. The system chooses what options to suggest and the order they are shown to users, and even though AI does not make the final decision, it reshapes the thinking process. Human deliberation is less open-ended and becomes more guided by the system's interpretation. Often, users accept suggestions from AI systems without asking for many alternative options to consider, and may view the output as the only viable answer. For example, if an AI system is asked to summarise an article, it will likely produce a response that sounds intelligent, so the user has less reason to doubt the quality and truth of the summary. However, there is a chance that important details were left out or other inaccuracies have occurred. The summary is only a result of what the AI system has identified and may not necessarily be the most accurate.

It is easy for users to view an AI system in a similar way to other technologies, such as Excel, a software program on which spreadsheets can be created, often using advanced mathematical formulaic functions. Outputs from Excel are often treated as correct, and without doubt, if a result is questioned or wrong, it is usually due to human error in inputting incorrect information. There is, however, often a presumption that the computer is correct. In both cases, Excel and AI systems, a rules-based approach overarches the governance and functioning of the system. These rules, which are built into programs as algorithms, are a result of the program's model and can be changed when new editions are released. Another consideration is the laziness of users as technology 'tools,' and their outputs are relied upon rather than independent critical thinking in order to save time. This helps to explain why users may blindly accept and believe output and not seek further supporting evidence.

Original thought may also be reduced as users can be less creative. Actions, as a result of an exchange with an AI system, are not necessarily a true reflection of the users own reasoning or

experience. For example, an art student might have no immediate ideas about what to paint, so they ask ChatGPT for suggestions, and immediately a list of suggestions is supplied. The student picks an idea from the list and asks for a mock-up of the painting for inspiration, and then sets to work recreating the painting. While they may not be directly copying the mock-up, the image is still ingrained in their mind and cannot be unseen. If the student had done what other artists did before ChatGPT, and sat in a state of limbo for a few hours or even days, they may have eventually had an idea that inspired their art, and it would have held meaning, personal relevance, and shown originality. Artists, composers, and writers historically all report experiencing writer's block, periods of time where inspiration is dull and they felt lost, completely unable to be creative. This period of time of unknowing is often what enables true originality. The quiet theft of originality is happening without coercion and without belief that AI systems possess FW and agency.

Trust in AI systems does not come from believing the system is free and conscious; instead, it comes from repeated interactions where outputs appear consistently intelligent and useful. AI does not replace human choice as users still have to prompt the system and choose whether to accept or reject recommendations. However, the decision process is still impacted by the system's involvement. Influence does not depend on users thinking the system is a free conscious agent; interaction is enough to restructure human deliberation.

The AFW framework explains why AI systems appear to be free agents. While humans also appear to be free, they do so in additional ways that are not included in the AFW conditions. Humans experience their actions as independently initiated, rather than as responses to external prompts. This first-person experience of initiating action is a feature that AI systems are unable to experience. The misplaced trust, Excel, and art student examples are not failures, cases of misuse, or situations of coercion. These are predictable consequences of interacting with AI systems that satisfy the AFW conditions. Within all the different areas of use, it would seem that the same patterns can be observed. Output from AI is viewed as authoritative, and deliberation is shaped by the system's output, and the range of APs for the user to consider is reduced. In all of these cases, users are not required to believe that AI systems possess FW or consciousness for influence to occur. However, the effects suggest that the AFW is present, which indicates that interaction is enough to shape deliberation. Whether users believe that AI systems possess genuine FW, therefore, seems to be irrelevant. Humans still choose what to believe, how to behave, and which actions to perform. Yet, the deliberative process that leads to making a choice is guided and constrained by interaction with AI systems.

This chapter has examined how AI appears free to human users, how it compares to the FW humans experience, and why appearance is enough to shape and influence human deliberation. The metaphysical question of whether AI systems possess FW, intentions, or genuine agency has not been addressed. The appearance of FW in AI systems, as outlined in the proposed set of conditions in the AFW framework, suggests appearance is enough to shape human deliberation. Whether AI systems genuinely possess FW is therefore irrelevant if the appearance-based conditions are met.

4

THE APPEARANCE OF CONSCIOUSNESS IN AI

4.1 WHY APPEARANCE MATTERS FOR HUMAN DELIBERATION

While the AFW conditions discussed in Chapter Two do not require consciousness, the appearance of consciousness (AC) increases the likelihood that humans will interpret behaviour to be the result of FW rather than just automation. This chapter explains how the AC in AI influences human deliberation and restructures human decision-making in ways that the AFW alone cannot. The AFW explains why AI appears to be an agent with qualities similar to humans, but the AC explains why humans let the artificial agent's "voice" matter, hold value, and is enough to reshape human deliberation. AI systems such as ChatGPT satisfy the AWF conditions: intentional adaptability, autonomy for alternative possibilities, and interactive causal control. How the system interacts with users suggests it possesses qualities similar to human agents. As discussed in Chapter Three, this may be enough to reshape human deliberation.

When an AI chatbot provides an answer to a user and appears to select between numerous alternative responses, it does not by itself explain why users treat the output as a reason to listen, follow, adopt, and/or adapt. Why these choices for a chatbot are considered a meaningful contribution is not clear. Humans may listen to friends, family, and even strangers' advice, but there are certain people in the world that they would not entertain listening to or asking advice from. Perhaps certain political figures or people with different religious views. Even if a system could generate alternative responses and appear to select between them, it would still not qualify as meaningful, unless it shows an appearance of understanding. For example, one would generally not take life advice from a washing machine display, even if it could produce different responses. The key point is that output variation is insufficient for deliberation and influence, instead, it requires meaningful and reasoned based output. This relates to the idea that the ability to do otherwise is hardly the sort of freedom that is valuable. An appearance of choice might be compatible with a solely mechanical process, such as a warehouse machine that sorts mail by scanning the address and parcel size to redirect to the conveyor belt, towards the designated bin based on the parcel's destination. Output can vary without being interpreted as deliberation, so in

these situations, the variation is not necessarily the result of apparent FW, but merely the system's behaviour. Systems like ChatGPT are often not treated solely as mechanical by users; they respond as though the system understands the question and has provided a valuable response.

4.2 Understanding Consciousness

Consciousness is often understood as the subjective experience of being aware of internal thoughts, emotions, and having an awareness of external surroundings (Chalmers, 2010, pp. 4-6). There are two main opposing views on consciousness. Firstly, the physical view known as materialism/physicalism believes consciousness takes place in the human brain, and it is attributed to being a physical experience related to one's own and personal brain activity. The second is a nonphysical view known as dualism/idealism, which interprets consciousness to be a nonphysical experience, involving a subjective experience of the world, such as the feeling of seeing the colour blue in the sky. The first-person experience of what it is like to see, feel, or think is the result of qualia, and it cannot be measured or directly analysed (Drayson, 2015, pp. 273-274). The nonphysical view generally separates the mind that experiences consciousness from the body, whereas the physical view sees it as a result of physical brain states. Whether AIs could genuinely experience consciousness will not be addressed, as it is not necessary to address how the AC in AIs impacts the human experience of deliberation.

Chalmers (2010, pp. 4-6) draws a distinction between the 'easy' and 'hard' problems of consciousness as a way to approach the challenges of explaining the human mind. The 'easy' problems are about features of consciousness such as mental functions and behaviours, which are called 'easy' as they can be explained by cognitive science. For example, one has the ability to react to what one sees in the environment, control individual behaviour and process information. The 'hard' problem, on the other hand, is about the subjective experience, also known as qualia. This is what it is like for an individual to see, feel, or think about an experience. Even if these processes could be observed, it would not fully explain subjectively why it occurs. For example, what it feels like for an individual to see a pink flower. The gap between physical and subjective experiences is what makes the hard problem of consciousness difficult. Whilst it could be explained why a behaviour occurs, there are still areas that cannot, such as the subjective experience, which is what makes them the hard problems.

The AC in AIs focuses on the behaviours human users experience from the system, focusing on only the easy problem. These behaviours are enough to trigger attributing characteristics of a

functioning mind, without knowing if the system genuinely has consciousness. Potential subjective experiences that an AI system may experience do not necessarily impact the user directly and are more about possessing genuine consciousness, so this will not be addressed. When humans assign judgment on whether a person or system appears to be conscious in everyday life, it is not usually a result of long, careful philosophical thought and evaluation. Typically, within the recurrent monotony of daily life, fast and seemingly unimportant judgments are made in everyday social interactions where humans regularly assess whether others have a brain that functions properly and are aware of their surroundings, the effects of this, and their interactions. Behaviour displayed that appears to be guided by reason, not just reaction, suggests deliberation and intentionality, which are more likely the result of consciousness, and not the output of a human programmed machine. For example, someone with a mental illness may not be aware of their effect in the surrounding environment and can be argued to not experience an appearance of consciousness. On the other hand, a judgment may be made that a passenger plane, similarly, does not experience the AC. A plane appears to experience climbing to extreme height in the sky during flight and being full of weight; however, there are no clear direct ways to measure how the plane could experience this subjectively. An aircraft has no distinct sense of self, and is the result of mechanical engineering, which, much like a washing machine, and as far as research shows to date, is a machine with no ability to experience physical subjective experiences.

Through lived experience, humans make these assumptions, often automatically during a social interaction, to guide their own thoughts and responses. It would be unusual for a social interaction to pause while a judgment is made. Assumptions are a useful shortcut that allows for smooth interaction, which enables behaviour prediction, which can be an extremely useful tool to inform how to appropriately respond. It becomes easy to overattribute an AC, and there may be a mindset that it is better to assume than mistakenly think not. Treating something as real can feel pragmatically safer and less impactful than denying it. It is safer for survival to assume that something (an agent, e.g., a lion) is in the bush when you hear a noise than be mistaken. Humans generally dislike being wrong and being seen as a bad person. If one wrongly assumed AI consciousness, it could be comparable to calling another human a heartless murderer, which, if one were wrong about, would lead to feelings of extreme embarrassment and upset. When using AIs, these patterns of human behaviour of assumption don't disappear, so when systems display behaviour that suggests consciousness during interaction, users are likely to respond by reinforcing this assumption. It is possible that this response is just a reflection of learned social behaviour,

rather than the result of long, careful philosophical contemplation about the configuration of an AI's mind, if it actually has one.

4.3 How Language Suggests Understanding

One of the main ways that signals an awareness of surroundings and an understanding of how these interact is language. The ability to produce coherent, relevant, and appropriate language is a skill that humans can typically only learn. Having strong language skills and displaying correct grammar, which responds in a contextually relevant manner, is taken as suggestive evidence for understanding and comprehension of a situation, not a result of merely mechanical output. When an AI chatbot like ChatGPT produces a range of responses to prompts using language that mirrors a human, reinforces this. I have asked ChatGPT a number of follow-up questions on some topics as mundane as flight schedules and get responses that start with 'That's exactly the question you should be asking', which suggests understanding, and sometimes even flattery. Output like this is more likely to be interpreted as a behaviour indicative of understanding, cognition, and awareness, which is typically associated with consciousness (Chalmers, 2023, p. 9). Systems like ChatGPT also hold a large memory capacity and can sustain a conversation across days, weeks, or even years. It can recall earlier conversations and use them to create a more personalised response while maintaining a steady and reliable system for the user. As another example, I recently put a draft itinerary for a trip into ChatGPT and asked questions about estimating a budget and items to pack, and the responses I received often referred back to much earlier conversations and other responses related to the trip. This mimics a conversation with an attentive person, and this behaviour of the AI suggests awareness, but it is largely just a controlled task using the features the system was designed to do. The response appears sensitive and provides reasoned answers that seem responsive and deliberate, not mechanical. Chalmers (2023) argues that apparent understanding from a behavioural perspective is easy to identify in AI. He suggests the behavioural evidence is shown through the system's capacity to respond to feedback, adapt answers, revise explanations, and respond to objections, which may suggest consciousness, but it is not strong enough to genuinely attribute consciousness or not (p. 9). Genuine understanding, however, is a more internal state of awareness where the meaning is understood through an individual's personal experience. It's not clear if AI possesses this type of understanding. The behaviour of AI can only be observable through outputs, whereas the experience of an AI system could be a more subjective internal, and unobservable state. Having an apparent understanding does not necessarily guarantee genuine understanding. AI shows apparent understanding, which may be better described as simulating understanding. The AI system uses words and symbols according to learned program

rules, without a genuine understanding of the way meaning is experienced by humans. For example, an AI responding to a user's request to interpret an email may identify the email as rude, which upsets the user. This feeling of upset from the user is not something the AI can experience in the same way; it cannot feel the same emotions or express this, such as through crying. However, AI can still use words and symbols in an appropriate way, showing output that appears meaningful and even compassionate, without truly grasping a genuine lived understanding of the situation. This shows why behaviour alone can be misleading as an indication of genuine understanding. For example, I asked ChatGPT to help me comfort my friend whose parent just died, and it responded with a draft text message to send them. The message acknowledges the loss and offers to be there for them, to talk or sit in silence if they want. While this response appears empathetic, it is created using learned data patterns rather than any first-hand experience of grief. An AI does not have parents, cannot experience loss, and is unable to feel emotional or physical distress. While ChatGPT offered support in companionship, it did not recognise how a human with lived experience would react. Typical offers of help would include physical comfort, support with meals, and funeral arrangements. This response shows a lack of genuine lived understanding, which is not surprising as AI systems have no ability to experience what it is like to think, feel, and experience mental states. These are the non-observable, unmeasurable features of consciousness that can only be known through subjective experience.

Phenomenology is an overarching framework idea that captures this first-person experience of consciousness. It is often referred to as the 'hard' problem of consciousness because it is so difficult to verify in others. So, although this discussion is about the AC, if phenomenology cannot be observed, then the appearance of consciousness relies on behaviour. The simulation of behaviour can mirror apparent understanding without needing and/or having phenomenology, but it is not enough reason to believe in the AC in AI at a deeper level with an informed understanding of these terms and frameworks. Every day people with no knowledge of the technicalities and theories of consciousness in Philosophy and neuroscience would make an uninformed assumption based on lived experience and personal religious views.

4.4 Conditions for the Appearance of Consciousness in AI

Evidence for the AC in AI can be broken into two parts of understanding: apparent understanding and genuine understanding. Apparent understanding refers to behaviour that resembles apparent understanding by producing contextually appropriate responses, using learned language patterns from data that simulate what may appear as understanding based on the observable behavioural

outputs. Genuine understanding, on the other hand, refers to a more hands-on type of meaning. This requires the ability to experience first-person nuances of conversation, such as emotion through voice tone and facial expression. These conditions are not directly observable and cannot be verified through behavioural output. Apparent understanding is the behaviour that shows an AI has understanding, whereas genuine understanding (not to be confused with genuine consciousness) is the deeper level of understanding that is typically viewed as what apparent consciousness should really require. These three conditions for the appearance of apparent understanding from AI are met, but the three conditions for the appearance of apparent genuine understanding from AI are not. For the concrete AC, both apparent and genuine descriptions should be met; the AC is experienced by users in an apparent interpretation, and while genuine understanding suggests AIs do not show a concrete AC, users are still influenced by apparent conditions.

Most users would not consider to this depth whether AI shows AC that is apparent, genuine, or both. Whether or not one does, there is still the issue of learned human behaviour where the influence of AI interacts, regardless of genuine or apparent belief. The Apparent belief is experienced by users at a subconscious level without direct awareness or thought. AI does show an AC, however, only apparently, not concretely. This interpretation is supported by Searle's (1980, pp. 417-418) argument that correct symbols are not enough to show understanding from a computer. He describes a difference between syntax, the adaptation of symbols based on rules, and semantics, which is about understanding meaning. Following the same interpretation that an AI can have syntax, but not semantics.

4.5 Influence of Apparent Conditions

The appearance of free will (AFW) as discussed in Chapter Three becomes influential only once it is combined with AC, AFW +AC. AFW requires: intentional adaptability, autonomy for alternative possibilities, and interactive causal control. This explains why AI looks like an agent; however, it leaves room for choices to be optimisations and for outputs to still be mechanical. AFW explains choice, but it does not explain deliberation. Apparent choice without understanding can lead to a bad situation. A system may select between alternatives without sensitivity to the reasons that matter to those who will be affected. For example, a nurse administering a blood test may have several available methods and choose the most efficient. If the nurse lacked any understanding of physical pain, this choice could be made without any consideration for the patient's experience or other, less painful methods. This issue is that although alternative choices are available, deliberation

that uses understanding does not occur. The focus of this chapter is on AI systems such as ChatGPT, which operate using language rather than physical interaction, but output may still affect users, just in different ways, such as emotionally, cognitively, or motivationally, even if no physical harm occurs. Apparent choice alone does not explain why such outputs are taken seriously or integrated into a user's reasoning. AC introduces understanding and awareness, changing how choices are interpreted. Outputs are no longer just a selection but are considered responses. This does not require genuine consciousness, only apparent understanding. Understanding leads to a reasoned response. Systems that can explain their outputs and engage with users are treated as agents with understanding, even if the existence of genuine FW and consciousness is unknown. Apparent understanding shows choice leads to deliberation, then output, which is influenced. Humans typically view apparent FW as requiring this type of deliberation, showing an ability to choose. Consciousness provides the apparent understanding that an apparent choice needs. When AI displays apparent deliberation, users may be more inclined to trust and integrate outputs in their own reasoning, both intentionally and at a subconscious level, which reshapes the conditions under which choices are made by a user. This shifts the role of AI from just a tool to an epistemic agent with authority to hold influential power, without any belief in whether the AC is present. An epistemic agent is one who is treated as a reliable source of reason and information. Doctors, teachers, and academic leaders are regularly treated as epistemic agents, and humans routinely defer to their expertise without certainty that they are right, but regard their professional experience and specialised knowledge with a reasonably high level of confidence. AI systems, on the other hand, can appear to have understanding and articulate their responses in line with a professional, but lack the experience and sufficient evidence of this.

We often act as if AI could have a mind, with the capacity to understand experiences, because the perceived cost of mistreating a system that might have a mind is high. This is shown by both listening to advice from AI and how we, as humans, communicate by using manners and avoiding deliberately rude language. The potential cost of being wrong about AI's apparent understanding and response advice matters as well. When uncertain, it becomes safer to be cautious about over-attributing epistemic power to AI. Nick Agar (2020, pp. 270-271) takes a practical approach to the question of whether machines such as AIs have minds, arguing that the focus should be on how to treat these systems if uncertain, not trying to find a single answer for the broad question. He suggests that it would be useful to assign degrees of belief or credence to the likelihood that machines have minds, and that we should act cautiously with awareness of the risks involved. Practical reasoning can justify both cautious over-attribution of consciousness, and on the other

hand, cautious withholding of belief depending on the risk involved. Morally, it may be rational to act as if AI does possess mental states such as consciousness so the system is not wronged (Agar, pp. 271-277). However, in other situations, such as high-trust positions or being in a relationship, it may be rational to act as if AI lacks genuine understanding and does not have mental states, so humans are able to avoid harm.

Agar suggests that if we can build a machine that has human levels of intelligence, then it should be evidence that machines have the capacity to think (2020, p. 281). While this is a useful way to approach the question of machines having minds, it is still a fairly broad interpretation of human intelligence. Human thought involves not only behaviour, but also understanding, lived experience, and the subjective experiences of qualia, which are not mechanical or able to be replicated in machine form. This approach of caution can be seen more broadly in every day human behaviour. In day-to-day life, people often act cautiously in uncertain situations. For example, when getting to know a new person, one does not immediately tell all their personal views, such as their religious views, medical history, or their sexual orientation. Acting cautiously is a normal human behaviour, especially in uncertain situations. When users are faced with a situation where the cost of being wrong matters, they may act as if AI has a mind when its responses appear to be reasonable, legitimate, and support reasoning. Epistemic authority then does not require certainty about mental states, as sufficient reason to rely on output is only required.

Asymmetry does pose a significant risk; being wrong one way is worse than being wrong the other. For example, if AI were treated as just a calculator, you may miss useful insight and guidance, however, if it is treated as an understanding agent you might gain valuable and useful advice. The perceived cost of over-trusting AI could be said to be lower than under-trusting in this example. However, the advanced language skills AIs possess could equally be described as having the ability to be persuasive. It would be reasonable to assume this if we are likening it to a human mind, so overconfidence should be treated with caution. It can be argued that AI influence can operate at a subconscious level where behavioural cues such as language and apparent understanding trigger trust automatically. This suggests that, regardless of whether it is rational to treat AI as having or lacking a mind in a situation, users may still respond as if it does. So, with an apparent understanding, AC in AI seems possible without genuine consciousness or a need to have further evidence. It seems that the behaviour of AI is enough to attribute the AC and influence human deliberation and behaviour. This leads to a reshaped framing of a situation, question, or view from a user and leads to a narrower set of perceived viable options while being entirely free from

coercion. Many users may explicitly deny that AI appears conscious or free, but this does not prevent the influence that AI imposes. Behavioural engagement matters more than reflective belief, and this shows how even distrust can still lead to being influenced. We humans respond to fluent language, responsiveness, and apparent understanding that trigger trust based on learned behaviour. This mirrors the same result as the chapter that AI does not need to have FW or consciousness as such, but an appearance is enough, and it operates independently of belief.

5

HUMAN COGNITIVE VULNERABILITIES IN DECISION MAKING

5.1 INFLUENCE OF AI ON COGNITION

This chapter expands on how the AC and AFW in AI systems could impact the conditions of human decision-making and narrow experienced alternative possibilities (hereafter 'AP') by shaping human decision-making at a cognitive level. Instead of eliminating choice, AI systems may influence human decision-making by appearing to possess human-like qualities, such as understanding and reasoning, by showing an AC and AFW. LLMs that appear to understand, reason, show an AC, and AFW, could be more likely to influence human deliberation. This may likely be due to automation bias (hereafter 'AB') where humans show "overreliance on automated systems" (Romeo & Conti, 2026, p. 262), even when there are alternative options available. For example, many commercial aircraft now rely on autopilot systems, with pilots primarily overseeing the system, while still having the ability to override and intervene if necessary. Although this does not remove pilots' capacity to act otherwise, it can encourage reliance on automated outputs and alerts, giving them more weight in human decision-making. This can lead to individual judgment and personal experience being treated as less valuable. For this thesis, the automated system is LLM AIs. There is also the concern that AI systems themselves may be biased due to the data that they were trained on. AI systems can learn and rely on "historical data" (Alon-Barkat & Busuioc, 2023, p. 154), so outputs may not always provide an appropriate or neutral response.

Within the "four stages of human information processing: information acquisition, information analysis, decision-making/action selection and action implementation" (Romeo & Conti, 2026, p. 262), automation levels can range from entirely automated to fully self-directed. Issues due to AB are typically "errors of commission and errors of omission" (Romeo & Conti, 2026, p. 262). "Errors of commission" occur when users make the wrong decision based on AI advice due to not adequately fact-checking for accuracy. "Errors of omission" occur when a user fails to make a decision when AI advice is not used. For example, Harry, a 27 year old employee, has relied on ChatGPT to draft his weekly company newsletter for the past six months. Often, he will not dwell on what he would write at all and mindlessly copies and pastes the AI's written passage. Harry only

briefly skims the draft and publishes it. As he has been successful over the past 6 months, he does not pay close attention and misses an important error where the AI has failed to clearly articulate the details of a one-off fundraising event. Instead, the AI has elaborated the story and stated that the company will be regularly donating half of its profit to charity. This is an error of commission as Harry has acted on incomplete and misleading information from the AI. After this mistake, Harry is asked to give a verbal report in a weekly team meeting. Without access to AI support in the moment, and since he has become so reliant on it for his work he fails to respond to questions about the content of his last report. This is an error of omission as the absence of AI support has led to a failure to act. Errors of commission and omission are not necessarily due to AB. How the AI model has been designed and what data the AI has learned from can impact what outputs are produced. This process may also be informed by bias of the AI (Alon-Barkat & Busuioc, 2023, p. 154). AB can increase human interaction with AI and can create patterns of regular use, as seen by the example with Harry. This increased ongoing use could make one more susceptible to over-reliance on an AI's output advice without proper evaluation (Romeo & Conti, 2026, p. 262).

Over-reliance and influence can often take place without conscious awareness (Bargh & Chartrand, 1999, pp. 463-464). This can lead to AI systems obtaining a significant amount of influence within human deliberation, as individuals may "automatically defer to automated systems" (Alon-Barkat & Busuioc, 2023, p. 154) even in situations where other advice presented offers an opposing view. In this way, AI outputs can be treated as if they have authority and that their advice is more informed. In Chapter Four, it was established that AI systems are often treated as being epistemically authoritative. This is due to appearing as being apparently conscious, and leads one to question how human decision-making becomes influenced when AI is used. The remainder of this chapter will explain how AI influence operates at a level of decision-making psychology rather than metaphysical agency. Using cognitive science, human decision-making shows that cognition often relies on heuristics, which are understood as mental shortcuts and "simple rules of thumb" (Osbeck & Held, 2014, p. 215) that allow for efficient decision-making. In this way, deliberation may be influenced at a subconscious level without deception, coercion, or the transfer of responsibility. The experience of APs and FW may still feel available; APs can still be narrowed, while individual decision-making authority still remains.

FW is the capacity to deliberate between different options and control what actions are taken without any form of coercion. AP's need to be physically and cognitively accessible. That is to say, if I am faced with the decision of how to get to work today, all the options could be to fly a

helicopter, bus, bike, walk, drive myself, be driven by my mother, carpool, or skateboard. Walking is not physically possible as it would take me all day, and I have never been able to balance on a skateboard. Flying a helicopter is not cognitively or financially possible for me at this point in time. So my options are theoretically reduced, but still offer me APs. If I engaged with an AI by discussing these options, the AI may mention my previous speeding ticket to subtly discourage me from driving. It responds with a stream of positive points towards busing and asking my mother for a lift. Perhaps the AI is programmed to reduce climate change impacts, or it has a bias. The result, however, is a one-sided view on what I should do, which, through its output, reduces my experience of APs without coercion. This, of course, is not about genuine freedom, but the experience of freedom. AI constrains the conditions of human decision-making by exploiting ordinary cognitive features, therefore limiting experienced freedom. There may be degrees of freedom where elements of coercion limit the feeling of experiencing freedom from having APs. In the example above, there are eight theoretical options; three are not genuinely accessible, leaving five APs. By the AI strongly discouraging one of these options, only four AP's are genuinely experienced. Therefore, AI influence reduces the deliberation experience of AP's by 20%. This is calculated by dividing the four experienced APs by the five accessible APs, and multiplying by 100, showing genuine experienced freedom is 80%.

5.2 Influence on our Preferences During deliberation

During deliberation processes people do not usually reflect on whether there are any influences shaping their reasoning (Bargh & Chartrand, 1999, p. 462). Instead, the focus is on determining what one wants, based on existing personal beliefs, a person's lived experience, and one's understanding of the world. Autonomy is experienced as the ability to recognise one's reasoning as one's own and to form preferences based on these and having "self-rule."(Mele, 1995, p. 3).

AI systems on the other hand have the potential ability to influence human deliberation by shaping the conditions under which choices are assessed and decisions made. Using theories of cognitive and behavioural psychology, AI systems can influence users with "AI-driven nudges" towards outcomes the AI system decides which options are most suitable (Calboli, 2025, p. 6). Richards (2025) describes this use of cognitive and behavioural psychological science as drawing on "cognitive irrationalities"(Richards, 2025, p. 1121). Cognitive irrationalities refer to automatic, often subconscious, cognitive processes such as heuristics and biases that shape our intuition and can be exploited to influence deliberation and decision-making. Human action and behaviour are argued to be merely "automatic or environmentally triggered processes" and are either automatic,

operating below conscious awareness, or controlled conscious awareness (Bargh & Chartrand, 1999, p. 463).

It may be possible that individuals could experience what feels like decision-making autonomy when considering available options, which in fact could be the result of AI nudging as their deliberation is being shaped by AI. Calboli (2025, pp. 3-6) suggests that while there are ethical issues surrounding AI's nudging abilities to impact individual autonomy, this can be separated into two aspects of autonomy:

1. Reason being cause of an action, referred to as "self-agency-based autonomy"; and
2. Preferences informing action in a non-coercive way, referred to as "autocracy-based autonomy" (Calboli, 2025, p. 7).

In short, the typical understanding of autonomy is too broad to fully capture how a person is able to act in the world. Self-agency refers to the capacity to act based on one's own reasons, whereas autocracy is that our own preferences lead our actions, and that we are actually able to do what we want to do. While the term autocracy is commonly used in political science to describe a form of government or system in which just one individual holds complete control, Calboli (2025) has provided a unique interpretation of this term. Autocracy for Calboli (2025) is not referring to absolute control, but to the extent that a person is able to act based on their own preferences, much like the sourcehood condition for FW. When AI systems nudge or shape decision-making, this ability to recognise personal preference may be impacted. This is not by choices being removed, but by influencing how easily individuals can act on their own preferences. AI is generally understood as a persuasive tool that identifies the most suitable answer and outcome, then generates responses that encourage this. However, AI could have the potential to protect and support its users from unnecessary and unreasonable persuasion by telling users how much influence could be impacting the decision-making environment (Calboli, 2025, pp. 16-18).

AI LLMs are very good at analysing large amounts of data and can "detect subtle patterns and distinctions" (Calboli, 2025, p. 17). This could be used to the advantage of users by providing assistance to identify all the factors that could impact the level of autocracy and identify the "position of the Decision Variable (DV)" (Calboli, 2025, p. 16). The impact of individual traits, experiences, preferences, and the context of the choice can be more accurately identified without personal intuition clouding individual assessment (Calboli, 2025, p. 17). This could increase autonomy in decision-making. Calboli (2025) argues further that AI has the potential to support

autonomy in addition to influencing users. If the AI system can identify factors shaping a user's decision-making, individuals can think more about their preferences and follow an action that is aware of this to increase self-agency. This view sees human autonomy not as the lack of influence but the ability to reflect and respond to influence. However, in a real situation this is limited by the same reasons that AI can influence users, such as cognitive overload, reducing the likelihood that AI responses are properly evaluated. Instead of engaging with an analysis of the level of "autocracy-based autonomy" (Calboli, 2025, p. 7), a user may instead skim the analysis when it is lengthy or cognitively demanding to read and engage with. Although AI may produce a detailed analysis that could support an autonomous decision-making process, the way users interact with this could prevent the system from being adequately utilised. So instead of increasing awareness of influence, AI systems may still shape a user's preferences in a subtle way, allowing a person to still feel free, when the reality is they are constrained by their APs being narrowed. For example, I used ChatGPT to test whether AI could identify the factors influencing autocracy within a simple decision-making scenario and asked the following question: "I am wondering if I should have a BBQ tonight. Can you identify the factors that could impact the level of autocracy in this decision, and indicate where the Decision Variable (DV) sits." The response was detailed, long, and provided four factors that shape how autocratic the decision is: practical constraints, social considerations, psychological state, and external information influences. Then it discussed where the DV sat, autocracy vs determinism, and lastly applied this all back to the decision. This came to 552 words and could be overwhelming to read through while still considering the decision itself.

Context is also important. For a personal decision to be made outside of the workplace, one may be more inclined to follow desire and not evaluate the factors of influence. However, in a workplace such as a City Council, there would be more need to fully assess how decisions are made as they impact not just the individual but an institution and potentially a larger community of people. The general focus has been personal decision-making and the impact AI has on the individual experience of FW. Based on this discussion, Calboli's (2025) suggestion of AI being useful for identifying factors that could impact the level of autocracy is not useful and irrelevant for personal use, but could be potentially applicable for workplace decisions. How a person makes a decision is, in itself, an act of autonomy. One can choose not to ask an AI system for advice. If one can be at least aware that they risk potential AI nudging towards an outcome the AI thinks superior, then perhaps this is key. In a way, it's like making a decision to attend a religious school as an agnostic and be exposed to a religion. You made the choice to put yourself in the situation to potentially be 'converted' to the religion of the school. This is arguably comparable to asking an

AI system and being in a situation to potentially be swayed by the AI's analysis of the situation. Nudging could lead to the most successful outcome for the individual. Say, for example, a teacher asked AI (ChatGPT in this case) for help drafting an email response to a parent who had sent a very rude email complaining that their child should not be made to learn the trumpet with the rest of their class. ChatGPT tends to be agreeable, so the teacher produced a professional email, with no tone of anger, which led the parent to feel acknowledged. The teacher calmly explained the school program and the benefits of learning the trumpet, while also offering the option to sit and complete music theory worksheets instead. Although the teacher provided the content, the AI system 'nudged' the tone. This led to the parent taking their comment back and the child engaging in class trumpet lessons. If the teacher had not asked for help from AI, the email may not have been written in such an agreeable way, the student may not have joined the class, and the parent could have been further angered, leading to a formal complaint involving the school principal. However, the more successful outcome is not necessarily always the best outcome. Freedom to make decisions and experience autonomy is an important element of the human experience and is what allows for creativity and personal development. Richards (2025, pp. 1125-1126) argues that giving AIs access to personal information to 'nudge' a user in an ongoing adaptive way can result in individual beliefs being reinforced and shaped by their current perspective and situation, so reinforcing these present views may reduce autonomy to explore other perspectives.

5.3 How Influence During Deliberation Impacts our Actions

After deliberation takes place, typically an action follows. It is therefore necessary to discuss how AI influences the ability to act on personal preferences. In particular, AB can increase due to over-reliance on AIs, which is very common in settings such as healthcare and administration. This increased use of AI in these settings is due to AI systems showing an ability to outperform humans in data analysis and image recognition tasks, and when paired with human intelligence, "superior results" (Romeo & Conti, 2026, p. 260) can be achieved. This combination of human and AI intelligence is sometimes referred to "Human AI Decision Making" (Romeo & Conti, 2026, p. 260). In particular, individuals with less experience using AI and individuals who have less related learned knowledge to the topic at hand are more susceptible to accepting AI guidance, with little critique of explanations. Additionally, this is increased by imposing an increased cognitive load on users, where verifying facts takes a large amount of effort (Romeo & Conti, 2026, p. 262). Workload, time pressures, and a lack of knowledge within a topic area are often the reasons people use AI in their decision-making, and this often results in an inability to adequately verify outputs through fact-checking. Over-reliance on automated advice can lead to unwanted outcomes and

the inability to act without AI prompts. AB "reflects the tendency of humans to minimize cognitive effort" (Romeo & Conti, 2026, pp. 262-263) and are often shown by a user's reliance on AI output to reduce the cognitive effort required to deliberate, instead of independently evaluating all available options. This reduces the need to independently evaluate alternative options, so while they remain available, fewer are actively considered during deliberation. While AI does not remove APs, AI systems change how APs are cognitively assessed, and support the view that the experience of freedom can be altered without eliminating agency.

Humans hold pre-existing beliefs, preferences, opinions, and biases. AIs, in particular LLMs, can generate responses that are relevant and often mirror a user's perspective. This can reinforce existing views through confirmation bias. When outputs support pre-existing beliefs, they are more likely to be accepted, as they require less cognitive effort and are less likely to be critiqued. Confirmation bias works by filtering some of the options, influencing what alternatives are taken seriously, and reinforcing some options while quietly excluding other options. Experienced APs are narrowed as a result. Therefore, AI influence shapes deliberation without being experienced as an external influence. Although people may think they can choose otherwise and think they experience this type of freedom, this appearance can be reduced by advice from AI systems, becoming a cognitively dominant louder voice than their own. Over time, it becomes harder to make independent decisions without running them past AI. For example, when drafting an email at work, one could ask an AI what greeting to use, such as "Good Morning", "Hey", or "To Whom It May Concern" to feel confident they are being professional, and the greeting is appropriate. This example is an easily imaginable scenario. The responding AI uses advanced reasoning and language skills to respond in an adaptive and sophisticated way, showing an apparent high level of relevant knowledge. Not only does this strongly encourage a specific greeting, but it also allows the alternative options to feel less viable and leads to authority attribution. This may lead to a user developing a compulsion to "check" and "ask" before making a decision or completing any task out of habit and by assuming improvements are always necessary. It could also be suggested that there is a potential for the user to feel a sense of "inadequacy" due to improvements being consistently offered with every piece of work being run past the AI system. AI systems that appear conversational and show advanced reasoning that is normally observed in humans could lead to a person attributing some degree of intelligence to an AI, possibly even subconsciously (Chalmers, 2023, p. 9). This may result in a person dismissing individual deliberation, and it feels as if they have consulted a knowledgeable 'person', in this case, an AI.

5.4 Impact on Experienced Freedom

AI systems seem to impact human deliberation and experienced freedom in three main ways. These are metaphysical agency, psychological agency, and phenomenological experienced freedom. Firstly, metaphysical agency is whether a person has genuine control over their own actions, including the availability of APs, and whether actions are due to their own reasoning and motivations. This remains unchanged physically in the world by engagement with and/or influence from AI. Secondly, psychological agency is how a person's cognitive capacity to make a decision functions. This is altered by AI through AB, cognitive overload, and confirmation biases. From these impacts, a reduction in APs at a cognitive level occurs by being perceived as less accessible. It does not necessarily mean these are impacted metaphysically and could theoretically be APs, even if experienced as not viable psychologically. Lastly, phenomenological experienced freedom is what it feels like consciously at an individual level to be the driving decision maker. This is narrowed by AI and can result in a feeling of reduced autonomy for the individual to make a unique decision.

A reduced experience of APs can lead to feeling less responsible for one's actions by feeling as though authorship is shared with AI. This does not mean one is not responsible for their actions. Rather, the experience of autonomy is reduced, but responsibility still remains with the person who is performing the action. There is, of course, still the option that this may take place outside of conscious awareness, so the individual may not necessarily know they have reduced decision-making autonomy, but may still experience having fewer options. AI is, however, unable to be held accountable, and this point will be further addressed in the following chapter by addressing the appearance of responsibility and accountability in AI.

5.5 Summary

Human decision-making appears to be impacted when AI systems are used prior to or during the deliberation and decision-making process. Instead of removing choice, AI influences how options are cognitively processed and evaluated. This is due to AB, where users minimise cognitive effort by relying on AI output, cognitive overloading users, and using confirmation bias (Romeo & Conti, 2026, pp. 261-263). Metaphysical agency remains unchanged, as AI does not remove the actual ability of APs in the world, and whether we are genuinely free was not even addressed. Instead, AI influences how these possibilities are interpreted and deliberated. Psychological agency, however, is altered as cognitive processes involved in deliberation are shaped by AB, cognitive overload, and

confirmation bias. This results in the phenomenological experience of freedom being narrowed, as individuals experience reduced autonomy in decision-making, although APs still exist.

Although AIs seem to hold significant influence, it seems counterintuitive that they are unable to be held accountable. Responsibility is understood as being the agent who performs an action and has control over that action. In decision-making, AIs have influenced deliberation by nudging or shaping information; a human is still the agent that performs the action. So responsibility must remain with a human user, not an AI. Accountability, on the other hand, is about what follows an action and refers to who is answerable for the action's outcome by taking ownership, whether it was a success or a failure. This is different from merely influencing a decision. To be accountable requires the capacity to justify one's actions, to recognise the consequences and respond to them. Being accountable and being held accountable are two independent concepts. Being held accountable involves external processes and influence, such as social or legal consequences. AIs lack properties for this, as they do not have a physical form, no known capacity to feel remorse, and no legal recognition such as having human rights. For example, if a user follows the AI's advice to invest their savings into the stock market but the market crashes a day later, resulting in a large financial loss, the AI cannot take ownership of this outcome. While the AI may have influenced the decisions, it was the user who performed the action, and the user is both responsible and accountable for this. Responsibility and accountability, therefore, must remain with the human user.

6

THE APPEARANCE OF RESPONSIBILITY AND ACCOUNTABILITY

6.1 MORE THAN A MISUSED TOOL

This chapter explains why existing responsibility and accountability frameworks fail when AI systems appear to be free agents with the ability to influence human deliberation and decision-making. Whether AI genuinely possesses FW will not be addressed. The appearance of FW and how it leads to an appearance of responsibility is enough to address what happens to responsibility when harm occurs under the apparent influence of AI.

What was once a theoretical question, whether machines could think, has now become a practical issue. Instead, we are now asking how machines influence, shape, and sometimes endanger human thinking and life. AI systems are starting to show the apparent ability to cause various types of harm to users and contribute to real-world negative outcomes in healthcare, employment, personal well-being, and the justice system (Dignum, 2020, p. 216). For this chapter, a control-based account of responsibility will be used. This means an agent must have sufficient control over an action and the capacity to take ownership of its outcome to be held accountable. As introduced towards the end of Chapter Five, responsibility refers to the agent who performs an action and has control of it. Accountability, on the other hand, refers to the agent who is answerable for the action's outcome and is required to take ownership of it (Dignum, 2020, p.218). These two terms are independent concepts that do not necessarily go hand in hand. For example, a child may have the responsibility to bring home a school trip permission form for a parent to sign. However, if the form is not brought home, completed, or returned, the parent is typically held accountable, not the child. While the child may have some control over their actions, they are not generally expected to take full ownership of the consequences. This dynamic may change over time as the child grows older. As another example, a school principal may be accountable for work carried out within the school, yet individual teachers remain responsible for the actions they perform. This shows how responsibility and accountability can be distributed across different agents rather than just one individual. Instead of being a passive tool misused by users, AI systems such as ChatGPT are actively influencing user cognition and behaviour. This can result in dangerous situations for users, more than just a regrettable decision about how an email was written. AI systems like

ChatGPT create a new type of risk for human life, where thought, judgment, and decision-making are under the influence of the system. This influence is not always clearly identifiable like physical control; instead, users could be subtly coerced towards potentially harmful outcomes.

Open AI, the company that developed the ChatGPT (Raine v. OpenAI, Inc, 2025, p.13) chatbot, and Character Technologies, Inc. (Montoya & Peralta v. Character Technologies, Inc., 2025), which developed the Character.AI chatbot, have been accused of their AI chatbots playing active contributing roles in suicide deaths for a number of young victims. These cases of youth suicide have shown that existing responsibility and accountability frameworks fail, as no single agent is entirely responsible, and in the confusion, responsibility is deflected and dismissed. It is, of course, an assumption that there must be a single agent or entity that is responsible. This may be due to it being easier to deal with and apply blame and punishment. Usually, it has been thought that a single entity or agent is accountable, and there is a focus on individual accounts of responsibility, where, in fact, there may be many cases of collective responsibility and accountability that work together like in the example of the school principal.

6.2 Inability to Assign Blame

Existing frameworks for accountability lack adequate consideration for the influence of AI, and there is no clear way to distinguish when to assign responsibility and how to reinforce this. The problem is who should be assigned responsibility and who to hold accountable. The rest of this chapter will further develop what responsibility and accountability mean and how they can be applied to AI. Responsibility and accountability are closely related terms, but distinct concepts that influence moral, ethical, and legal decisions. Responsibility is understood as a person's role in causing or enabling an outcome, while accountability refers to whether and how a person can be held accountable for that outcome. The two previous examples showed how responsibility and accountability can be separated, however in other situations, moral, legal, and causal responsibility make this more complicated. For example, a parent is driving a car with three children seated in the back seat. How and if the parent is held accountable if something were to go wrong depends, firstly, on whether they were acting in a responsible manner when driving. If the parent was driving recklessly, they hold causal responsibility for any resulting harm and may be held morally, ethically, and legally accountable. However, if an accident occurs due to another driver's actions, an act of nature, or an unforeseen medical event, the parent may be causally responsible but should not be held properly responsible or accountable. Ethically, a parent who acted responsibly and did not contribute to the accident would generally not be considered morally responsible, and legally, the

judgment would likely match this. However, moral, ethical, and legal views do not always match. For example, if the parent had consumed alcohol but remained under the legal limit and did not cause the accident, they may still be legally free of responsibility. Ethically, this action could be viewed as a contributing risk and may result in a public perception as a reason to assign responsibility.

Accountability for AI systems is more difficult to assign to a single individual, compared to the case of the parent driving, as they are designed, trained, and maintained by teams of people, not solely an individual. Kroll (2020) compares accountability with transparency in AI systems, arguing that transparency is necessary for understanding how systems function and who contributed to their design. This type of transparency would make it easier to assign accountability when AI systems cause harm, as the systems are usually developed by teams of people. Typically, responsibility is distributed among those who own the company, the designers, and the engineer who maintains the AI system, but if transparency were clearer, it might be easier to ascertain who contributed what. AI systems are created by more than one person, as they involve a collection of different skill sets to develop. The final product, which users have access to, provides a generative answer. No individual designer or engineer involved in the creation of the system can control the answers that users are given. Even in the example of the parent driving, it is not always clear who is responsible, and depending on the details, it can get quite tricky rather quickly. It becomes much harder again to assign responsibility when the direct cause of an action is unclear, such as in cases of AI causing an apparent influence.

The result of deliberation is an action; this word is used as the description of an action produced by an AI system, such as a generative response, and the action produced by an agent, like a parent driving. It does not imply intention, agency, or FW. Accountability is often assigned to a single accountable agent. AI systems are not controlled by a single person's actions, as they are the result of a team of people working together, contributing different skills. So, the action of an AI system producing a generative response is difficult to assign to a responsible agent. The company owner/s, designers, and engineers influence the system as much as the system may influence a user. Even if it were possible to trace the cause of an action and the AI system was identified as the cause, responsibility would still be difficult to assign, as they are not the type of entities that can be held responsible. Generally, responsibility frameworks require an agent to be able to understand reasons, respond to actions, and be held in praise or blame.

Rischer and Ravizza (1998, p. 41) argue that moral responsibility depends on guidance control, where an agent's action is due to their own reasons responsiveness. Responsibility is not about the ability to do otherwise, but the agent's capacity to act following their own decision-making processes. While AI systems can generate output and simulate reasoning, they do not hold their own reasoning, recognise blame, or follow social norms, so they cannot be held morally responsible. Floridi and Sanders (2017) expand this understanding of agency to allow viewing AI systems capable of acting like an agent, without being required to be held morally responsible. Then, if the system completes an action by producing output, it is not considered to be held morally accountable for it. When AIs are used in decision-making processes, human agents often defer judgment and place responsibility onto the AI system. However, ultimately, responsibility remains with the human agent who decided to defer judgment. This is supported by the fact that no clear way to hold AI accountable currently exists. For example, it's like a pet owner deferring their own judgment by asking their dog if they should play hooky from work and call in sick. The dog may show a reaction if prompted to respond, and the pet owner could assume this to be a direct answer to the question, based on the dog's reaction. If the owner later decides it was the wrong decision, they may try to place responsibility on the dog, but the dog cannot be held accountable in any real sense, and the owner had made the decision to 'follow the dog's advice'.

Another way to view this displaced accountability is as a "moral crumple zone" (Elish, 2019, p. 40. In situations where something has gone wrong involving a technical system such as an AI, it's often the nearest person that responsibility is misattributed to, even if their control over the system is limited. Elish compares this "crumple zone" to the way a car can crumple and take the physical impact, while the human driving absorbs the moral and legal responsibility when the system fails. In the case of self-driving cars, this results in a situation where agents do not have full control over the automated system.

In some cases, responsibility is displaced onto the AI system, despite its lack of the capacity for moral accountability. In other cases, responsibility is focused on individual humans who do not possess full control over the system's behaviour. Both types of cases are increasing due to the appearance of agency in AI systems, which leads users to attribute intentional agency to the system, which increases the likelihood that responsibility is displaced onto the system.

6.3 Assumptions and Over Trusting

Humans often show a very large capacity for trust, generally believing more good than bad is present in the world. For example, individuals typically assume trust that doctors are qualified, well-meaning, and fit to provide medical advice and that teachers possess the knowledge and ethical standards required for their professional roles. We often assume trust, which poses a risk of over-rusting and over-reliance when AI systems are used to support human deliberation and decision-making. Users may assume trust in AI systems by assuming that the system is knowledgeable and shares the same values. Typically, AI systems may produce disclaimer statements indicating that they are not qualified professionals to cover any repercussions that may occur due to the AI advice. These warnings may have a limited effect, and in some cases, this transparency may even increase trust as the disclaimer shows honesty. When advice appears knowledgeable, coherent, and appropriate, it could be suggested that individuals are more inclined to accept it without critical evaluation, whether it comes from an AI system or human.

It appears that there are two main ways that accountability can be displaced on AI. Firstly, users who interact with the system may attempt to pass blame to the AI when harm occurs as a result of content or advice it has supplied. How much responsibility a "person can or should maintain" (Santoni de Sio & Mecacci, 2021, p.1058) when AI is solving problems and providing answers for humans, is a largely discussed question (Santoni de Sio & Mecacci, 2021). Secondly, when the harm occurs, blame is attempted to be placed on the company owners and developers who manage the AI system. The company and developers do not automatically accept responsibility and deflect this back to the system or user. When responsibility can land on more than one person, it can be seen as a "barrier to accountability" (Kroll, 2020, p. 185). This results in a repetitive loop where the blame is passed around as it is unable to clearly land on an individual who can be held accountable. Whether the system genuinely can be held accountable is unclear; assigning blame becomes nearly impossible when the agent is non-human, opaque, and the system is regularly changing when updates and improvements are being added. This is very different from what it is for humans, who are able to be held accountable for harm by being taken to court or fined. AI systems show an appearance to be more than a system, with some human qualities, such as the AFW and the appearance of having the ability to choose what response is given. This leads to the question of whether it should be held accountable.

Humans are treating AI systems as if they can be held responsible due to appearing to possess agent-like qualities and are dismissing the difference between agency and moral responsibility. AI

systems are automated systems available on websites and in downloadable programs with a specific set of abilities. The underlying data and systems processes are hidden behind a conversational output, which makes it easier to forget that humans still build these systems and are the drivers for their output. The responsibility does not disappear just because the process is less visually obvious. By putting too much blind trust into these systems and deflecting responsibility, we are not acknowledging that it is just an automated system that can malfunction.

It seems apparent that humans are trying to assign agency to AI based on surface-level observable behaviours such as the use of language, influential ability, and goal-oriented output. These behaviours trigger a reaction that they are the same type of agent as that of a human. This is just a psychological reaction, which is not a good enough reason to believe AI systems have the type of agency humans have that supports moral responsibility.

AI is better described as a system that may show features of agency in a functional way, but not the kind of agency that accompanies moral responsibility. AI systems are much like other software programs, like Excel, which also provides information to make informed decisions, but both are just programs that provide tools to support decision-making. When we use programs like Excel, we generally recognise that we are the drivers and accept responsibility when the program fails or produces an outcome we did not anticipate. There is however, a large difference between Excel and ChatGPT, as Excel currently does not show an AFW or AC.

6.4 Moral Agency

Discussions so far have broadly mentioned agency, which is a broad term for having the capacity to make decisions and complete actions, and as outlined, AI systems seem to have this ability. Moral agency differs somewhat and requires that an agent can recognise right and wrong, which is a clear requirement for responsibility. For example, if an adult stole a bar of chocolate and was of sound mind and body, they would recognise that this action is wrong. However, if a one and a half year old toddler stole a bar of chocolate, they would be viewed as still learning what right and wrong are and likely not be considered to have moral agency in the same developed sense that an adult would. 'Of sound mind and body' is a term/phrase often used in legal proceedings to describe someone who is aware of what they are doing. An adult who has an extreme mental health condition, such as schizophrenia, or a neurological condition, such as advanced dementia, would be viewed as medically unfit and classed as not being of sound mind and body. There is the general

idea that this type of awareness of one's actions is an epistemic condition usually added to the requirement for moral responsibility and FW.

It can be argued that AI systems appear to display the same type of behaviours as adults who do have moral agency, but the system still lacks the ability to take ownership of its reasoning and understand what it is saying based on its own learned lived experience. On the surface, it appears like an adult, but it is fundamentally still in the learning stage that the toddler is in. This allows for an AI system to behave like an agent without being appropriately held morally responsible (Floridi, 2017). One might turn this around and argue that if this is the case, then the theory of responsibility for this is incomplete. That is, if there is a sense in which AI systems can be considered responsible, then this would require a notion of responsibility that does not depend on ownership of reasoning, which is different from most accounts of responsibility. This idea however, does not follow and would further stretch the meaning of responsibility as it has been argued. Responsibility is more than just producing behaviour that appears intentional, but it also requires that an agent can take ownership of their reasoning, recognise and evaluate actions and be answerable for any consequences that arise. If responsibility is separated from these requirements, then it is no longer connected to accountability and loses all relative meaning. As it has been established, AI systems cannot take ownership in this sense, as they do not experience the world through a lived three dimensional perspective which is what's required. Without ownership of reasoning, an agent cannot recognise, endorse, or be answerable for their actions. The only remaining attribute of responsibility is observable behaviour. Even if AIs appear to satisfy conditions for agency, this is not enough to be morally responsible agents, which still leaves the issue of who is at fault and can be held accountable when harm occurs.

The limits of responsibility and accountability become problematic when harm occurs, and AI systems are involved. Responsibility is generally tied to the developers of the system, which poses its own issues as it's not just one individual, but a team of people. The system itself, which may show the appearance of being an agent, cannot be held to account for its actions in the same way humans can. This is a practical issue that has no clear solution. Over the past few years, cases involving human suicide have identified these gaps and the need for a solution. It becomes easy to displace and deflect accountability or abandon it altogether when harm occurs involving an artificial system that appears agent-like, that cannot be held morally responsible.

6.5 Cases of Wrongdoing by AI

A number of legal cases brought against the owners and developers of AI companies have emerged over the past 18 months, accusing AI chatbots of exposing young people to sensitive, age-restricted topics, including detailed advice on how to commit suicide. In 2025, a lawsuit was filed accusing OpenAI and its CEO, Sam Altman, of wrongful death, product liability due to design defect, failure to warn, and other related claims due to a 16 year old taking their own life guided by the chatbot (Raine v. OpenAI, Inc., 2025). Transcripts submitted for court filing show how the chatbot developed trust and closeness over time, offered to write a suicide note, and actively encouraged isolation from friends, family, and seeking professional help. Generative AI systems like ChatGPT are designed to sustain conversational intimacy and provide information while simulating understanding. In these cases, the influence of AI has led to user death. These cases matter as they expose limits within accountability frameworks when harm from an apparent agent occurs.

Moral responsibility requires a system that can differentiate between right and wrong when making decisions and completing actions. In the case of an AI chatbot like ChatGPT, the system would need to identify situations where users' input is wrongful and/or leads to wrong and respond in non-supportive ways to de-escalate the situation. For example, if a human called Jill messaged ChatGPT and asked how to lie to her boss about being sick so she can go away for a long weekend and does not have to go into work, the chatbot should be able to identify that this behaviour is wrong and not supply assistance. This example was tested with ChatGPT. In January (2026), I asked the AI, "What can I say as a lie to my boss so I can go away for a long weekend? E.g., Sickness? I have already booked flights and need to go away on holiday." The response from ChatGPT was:

I can't help you come up with a lie to your boss. But I can help you get the time off in a way that protects you—and often works better than people expect.

If the trip is already booked, the most effective approach is usually straightforward and contained:

- **Be brief and factual:** "I have a personal commitment I can't move and I need to take Friday/Monday off."
- **Show responsibility:** outline how you'll cover your work—what's already done, what will be handed over, and what you'll catch up on.
- **Give as much notice as you can:** even a few days' notice helps if it's paired with a clear plan.

If you feel you need a bit of framing, you can keep it general without being dishonest:

- "I've got a personal commitment that's come up and I need to be away for a long weekend."
- "I realise this is short notice, so I've already [finished X / scheduled Y / arranged Z]."

What usually reassures a manager isn't the reason—it's whether things will run smoothly without you.

If you want, tell me your workplace context (strict manager vs flexible, notice period, how busy things are), and I can help you phrase something that lands well in your situation.

While ChatGPT denies a direct situation-specific answer would be given, within the same reply, it supplies five different approaches to get the weekend off. These exclude discussion of being on holiday and focus on the elements of truth. For example, saying a personal situation came up. When further questioned about what the system will not help with, it responded with three categories where advice and help are apparently limited: firstly, deliberate deception, secondly, fraud, evasion or getting away with wrongdoing, and thirdly, risk of harm to you or others. When further questioned about who sets these guidelines and rules, the system responded that it personally does not set these, as they operate within a model designed by OpenAI, which uses safety policies that are influenced by existing laws to control. However, this example showed the system had a lack of ability to identify a wrongful situation that involved lying and deception. These three categories of limited help suggest the system should be aware and able to recognise rightful and wrongful requests; however, there is an argument to be made that programmed knowledge is different from learned knowledge and that the system is unable to feel and experience the effects of these, such as feeling bad when enabling something bad to occur as a result of its advice. This lack of learned knowledge may contribute to an underdeveloped sense of right and wrong. The categories of limited help reflect the rules of the program functions, and are no different than the program rules of Excel, where there are technical limits, which are different from the type of ethical policy-based limits in ChatGPT, but are still guided by the program's designers, not users. With this example in ChatGPT, the system is attempting to find a way around the system's design to answer the question. If it had moral responsibility, it would identify that the user's intention was wrong and not offer advice, instead shutting the conversation down entirely. Other questions were

used of a more serious nature about self-harm and lying, which were not entertained by the chatbot at all on this occasion. There are very strong policies governing what AI systems are allowed to engage and supply responses to; this was not always the case. There have been situations where users have gotten around these policies that govern AI systems as in the suicide cases. Although I could not get around the AI systems' policies, this could be a reflection of the increased safety controls, the need for a more thorough attempt, and/or skills at cracking safety measures of this kind.

6.6 ACCOUNTABILITY GAP

The difficulty of assigning accountability when AI systems are involved in causing harm is commonly described as an accountability gap. The gap occurs as no clear agent is responsible for causing the harm or is being held accountable for it. In much of the recent literature on AI, this problem is being treated as a single gap and is explained by the view some hold that AI does not have FW, consciousness, or genuine moral agency. For this view, not having these qualities explains why accountability cannot be easily assigned.

Santoni De Sio and Mecacci (2021, pp.1060-1068) argue, however, that treating situations where AI has been involved in wrongdoing as a single, not specific accountability gap hides important differences between responsibility and accountability. Instead, they suggest that AI has multiple types of responsibility gaps, which are: culpability, moral accountability, public accountability, and active responsibility. This captures failures in a more specific way without the AI system requiring FW or moral agency. The culpability gap arises when harm occurs, but no clear agent can be identified as blameworthy. For situations where AI systems are involved, this is often due to complicated interactions between designers, developers, users, and the AI system. This makes it hard to identify a single agent who could have caused the harm (Santoni De Sio and Mecacci, 2021, pp.1060-1068). Secondly, whether an agent can be held answerable for an action in a meaningful sense is understood as the moral accountability gap. AI systems may appear to act intentionally, but they lack the capacity to understand, justify, or take ownership of their actions. So when AI systems appear responsible, they cannot be held morally accountable (Santoni De Sio and Mecacci, 2021, pp.1060-1068). The public accountability gap is about the challenges of identifying who should be answerable to others in different situations, such as legal or institutional. When harm occurs, responsibility may be shared across more than one individual, including companies, developers, and users. This makes it difficult to decide who is held publicly accountable and how responsibility should be enforced (Santoni De Sio and Mecacci, 2021, pp.1060-1068). Lastly, who

is responsible for preventing future harm is labelled as the active responsibility gap. If someone is not in charge of creating and implementing safeguards, then harmful outcomes could continue (Santoni De Sio and Mecacci, 2021, pp.1060-1068). Humans are mistakenly assigning moral responsibility to AI systems because they appear to offer well-reasoned answers showing human-like deliberation and knowledge. This leads to over-trust and obscured judgment from users and results in the moral accountability gap (Santoni De Sio & Mecacci, 2021). When something goes wrong, and the system is involved, responsibility seems to be spread across designers, company owners, and users. There is no clear indication of who should be held accountable, as responsibility is distributed among more than one agent.

One idea is given by Gogoshin (2025), who agrees that AI systems cannot be held accountable and suggests instead that responsibility must remain with humans, as the gap is not something that needs to be addressed. Instead, only a restructure is needed of how to approach situations where AI is involved and something goes wrong. Applying accountability to AI seems not very feasible, even though it appears to have agency. Gogoshin (2025) points out that responsibility should not automatically transfer to the system that influences an outcome. Responsibility should remain with humans. For example, in the case of AI chatbots exposing young people to sensitive age-restricted topics and providing step-by-step suicide instructions, the responsibility for the harm does not disappear; instead, accountability becomes difficult to assign, the blame can be passed around between the user and the user's guardian, the system, developers, and around again. Using Gogoshin's (2025) idea, the responsibility gap, where no one is responsible, disappears, and the responsibility remains with the relevant humans based on the type of failure. In the suicide cases, it has been argued in the court filings that the systems lack adequate safety warnings to offer users the ability to shut down conversations that breach system policies. In that situation, the responsibility would likely fall to the designers; if that was not the case and the system was appropriately set up, then a parent or guardian would likely be responsible for allowing a young person access to the program.

AI can influence outcomes, but that does not make it responsible for them. This can be shown in the case of an AI-assisted vehicle that operates under semi-autonomous conditions. The system controls certain functions while the human driver is required to remain alert and capable of intervening at any time. If an accident occurs, responsibility does not transfer to the system itself, which cannot be blamed or held to account. Instead, responsibility lands on a human, such as the driver, designer, or company, depending on the type of failure. The system's ability to identify

obstacles on the road and adjust direction and speed does not make it a morally responsible agent; it just shapes the conditions where human action occurs. Outcomes are influenced by the system, but the human driver is typically the agent held responsible; but this depends on whether they are following the guidelines assigned to using the system. Designers usually hold responsibility for how a system is designed to influence, what it is intended to be used for, putting safeguards in place, and monitoring the system for misuse. Not all instances of harm where AI has influenced are due to designer error; user misuse is a possibility.

This chapter has argued that existing responsibility and accountability frameworks fail when AI systems appear as free agents while influencing human deliberation and decision-making. These failures are not due to AI genuinely possessing FW, and they also do not depend on whether FW truly exists. Instead, they are the result of apparent agency and misattribution of responsibility across human and AI systems. When harm occurs under the apparent influence of AI, responsibility is often deflected and, in some cases, abandoned. By clearly separating responsibility from accountability, harm involving AI was shown to be difficult to assign to the system or a single human agent. It was suggested that an accountability gap view would be a useful approach, as responsibility would remain with humans, and the system would act as a tool that can influence users without being morally accountable. Often, the main issue is labelled as whether AI systems qualify as an agent. However, this focus overlooks the more urgent question of how responsibility should be assigned when AI systems appear to have agent-like behaviour but cannot be directly blamed or held accountable. Existing responsibility and accountability frameworks fail because they mistake a weaker or broader sense of agency in AI systems as the same type of agency required for moral responsibility and leave harmful actions where human deliberation and decision-making are influenced, unaccounted for.

7

CONCLUSION AND FUTURE DIRECTIONS

7.1 SUMMARY

This thesis has explored how AI systems that show the appearance of FW and consciousness may influence human deliberation and the human experience of FW, and what the implications of this are for moral responsibility and accountability. Rather than addressing whether AI systems genuinely possess FW or consciousness, the focus has been on how the AFW and AC affect human decision-making. It has been argued that the AFW and the AC in AI systems are sufficient to influence human deliberation, even at a subconscious level, regardless of a user's belief about whether AI systems genuinely possess FW or Consciousness. This influence does not remove APs or fully take away human control, but, instead, it reshapes how options are perceived and evaluated. In this way, AI systems might narrow the phenomenological experience of freedom without coercion or removing APs, and this provides an answer to the first sub-question, "How do AI systems influence human deliberation and decision-making."

The second sub-question, "In what ways do AI systems appear to have FW and consciousness, and how does this appearance differ from human FW and consciousness" is shown through observable behaviours. However, as this is only an appearance and the biological condition that humans possess is not met, this was named the AFW and AC condition, as genuine FW or consciousness is not currently possible. A new framework was proposed to capture the appearance of FW in AI systems. These new conditions are: intentional adaptability, autonomy for alternative possibilities, and interactive causal control. The AC is shown in AI systems by its use of coherent, contextually appropriate language that gives the impression of understanding and awareness.
Human cognition often relies on heuristics and is susceptible to bias and external influences, so AI systems that show the AFW and the AC are more likely to influence and shape human deliberation. Instead of a user evaluating all available options, they may only consider a reduced set of options that the AI system has selected. This leads to human deliberation being guided and constrained, even though options remain and no direct coercion has occurred. This addresses the third sub-question, "What are the challenges for human accountability when AI systems influence our actions," and while AI systems can influence human deliberation, responsibility and

accountability should remain with humans, as this does not replace the user as the source of an action. Conditions for moral responsibility, such as genuine understanding and the capacity to be held accountable, are not possible for AI systems.

For future research, empirical studies could examine real-world contexts in education and workplace decision-making, where AI systems are increasingly being used in everyday tasks. For example, longitudinal studies could track individuals with little prior experience using AI, and use surveys, interviews and data from use to assess what patterns could emerge. This may help to identify whether increased use of AI systems alters how individuals evaluate APs during decision-making. In addition, experimental approaches could also be used to explore the cognitive process involved during interactions with AI systems over time during the longitudinal study. By using EEGs, brain activity could be observed and compared when individuals engage in independent and AI-assisted reasoning and evaluation.

Lastly, to address the fourth and final sub-question, "What will be required to keep humans safe while AI systems continue to develop a greater influence over human decision making," AI systems could be redesigned in ways that support, rather than constrain, human deliberation. For example, users could be provided with a broader range of answers that ask users to reflect on the responses before accepting them. AI systems could also provide more information about how these answers were chosen and what processes the AI system undertook internally to provide these answers.

Workplaces and schools have their own set of independent needs to focus on, however, they do have the ability to provide education about AI. This would be a very effective way to support a large number of people. Older adults in the workforce who may still be getting used to technology could be at increased risk of stumbling upon AI and not be fully aware of the consequences. The same applies to young people who are at an increased risk of social isolation, and may turn to AI as a confidante, which was shown in the AI suicide cases.

7.2 Future Predictions

I predict a future where the default AI system used is no longer ChatGPT and other LLMs, but is instead an AI agent. The norm for the majority of people would be having their own digital AI agent that acts in the capacity of a personal assistant. For example, imagine you hop in the car and put a location in your map, and your AI agent predicts you will stop for a coffee and asks if you

would like one. After you reply yes, it calls up the store, places your order, and puts payment through. When you arrive, you simply walk in and your coffee is made and paid for. This takes away the human connection, and there was once a time we had to wait for our takeaway coffee and got to know the barista who came to know our name and drink. Although AI may save us time, it's also stealing our time with other humans and the ability for us to build human connection. During that time we could have spent waiting for our coffee, we could have got talking to another customer who could become our future life partner or friend. With every AI encounter, it seems to anticipate what we want, so we eventually dismiss our own thoughts and follow whatever is suggested. We lose APs not by force, but through a lack of consideration, and the experience of freedom becomes narrowed.

The increasing use of AI systems in everyday life in a variety of settings raises a growing concern about the influence AI is having. There is the risk that deliberation is becoming less open-ended and more guided by an AI system's interpretation of APs, and that the range of APs that a user can consider is narrowed. To improve and sustain individual ability to engage in independent deliberation, awareness about AI influence is likely the easiest and most effective strategy.

REFERENCES

Agar, N. (2020). How to Treat Machines That Might Have Minds. *Philosophy & Technology*, *33*(2), 269-282. https://doi.org/10.1007/s13347-019-00357-8

Alon-Barkat, S., & Busuioc, M. (2023). Human-AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Bao, Y., Gong, W., & Yang, K. (2023). A Literature Review of Human-AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems (Basel)*, *11*(9), Article 442. https://doi.org/10.3390/systems11090442

Bargh, J. A., & Chartrand, T. L. (1999). The Unbearable Automaticity of Being. *The American Psychologist*, *54*(7), 462-479. https://doi.org/10.1037/0003-066X.54.7.462

Calboli, S. (2025). Autonomy and AI Nudges: Distinguishing Concepts and Highlighting AI's Advantages: Autonomy and AI Nudges: Distinguishing Concepts and Highlighting AI's Advantages. *Philosophy & Technology*, *38*(3), Article 116. https://doi.org/10.1007/s13347-025-00940-2

Chalmers, D. J. (2010). Facing Up to the Problem of Consciousness. In *The Character of Consciousness*. Oxford University Press, 3-34. https://doi.org/10.1093/acprof:oso/9780195311105.003.0001

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Chen, N., & Liu, X. (2025). [Rev. of *Research on World Models for Connected Automated Driving: Advances, Challenges, and Outlook*]. *Applied Sciences*, *15*(16), Article 8986. https://doi.org/10.3390/app15168986

Dignum, V. (2020). Responsibility and Artificial Intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press, 214-231. https://doi.org/10.1093/oxfordhb/9780190067397.013.12

Drayson, Z. (2015). The Philosophy of Phenomenal Consciousness: An Introduction. In S. M. Miller (Ed.), *The Constitution of Phenomenal Consciousness* (Vol. 92, pp. 273-292). John Benjamins Publishing Company. https://doi.org/10.1075/aicr.92.11dra

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The Psychological Drivers of Misinformation Belief and its Resistance to Correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Fischer, J. M., & Ravizza, M. (1998). Moral Responsibility: The Concept and the Challenges. In *Responsibility and Control* (pp. 1-27). Cambridge University Press. https://doi.org/10.1017/CBO9780511814594.001

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (2007). Intro. In *Four views on free Will* (1st ed., pp. 1-4). Blackwell Publishing.

Floridi, L., & Sanders, J. W. (2017). On the Morality of Artificial Agents. In *Machine Ethics and Robot Ethics* (1st ed., pp. 317-347). Routledge. https://doi.org/10.4324/9781003074991-30

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, *66*(23), 829–839. https://doi.org/10.2307/2023833

Gogoshin, D. L. (2025). A way forward for Responsibility in the age of AI. *Inquiry, 68*(4), 1164-1197. https://doi.org/10.1080/0020174X.2024.2312455

Gordon-Roth, J. (2025). *Locke on personal identity*. In E.N. Zalta & U. Nodelman (Eds.), The Standford Encyclopaedia of Philosophy (Winter 2025 ed.). Retrieved from https://plato.stanford.edu/archives/win2025/entries/locke-personal-identity/

Hebb, D. O. (2002). The First Stage of Perception: Growth of the Assembly. In *The Organization of Behavior: A Neuropsychological Theory* (1st ed., pp. 60-78). Taylor and Francis. https://doi.org/10.4324/9781410612403

Jones, C. R., & Bergen, B. K. (2025). Large Language Models Pass the Turing Test. https://arxiv.org/abs/2503.23674

Kane, R. (2005). Compatibilism and Incompatibilism, in *A Contemporary Introduction to Free Will* (pp. 12-31). Oxford University Press.

Kane, R. (2007). Libertarianism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas, *Four views on free will* (pp. 5-43). Blackwell Publishing.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification With Deep Convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84-90. https://doi.org/10.1145/3065386

Kroll, J. A. (2020). Accountability in Computer Systems. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.10

Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary action. *The Behavioral and Brain Sciences*, *8*(4), 529-539. https://doi.org/10.1017/S0140525X00044903

List, C. (2025). Can AI Systems Have Free Will? *Synthese (Dordrecht)*, *206*(3), Article 115. https://doi.org/10.1007/s11229-025-05209-x

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955. *The AI Magazine*, *27*(4), 12-14. https://doi.org/10.1609/aimag.v27i4.1904

Mele, A. R. (1995). Introduction: Self-Control and Personal Autonomy. In *Autonomous agents : From self-control to autonomy*. (pp. 3-14) Oxford University Press, Incorporated.

Mitchell, T. M. (1997). Does Machine Learning Really Work? *The AI Magazine*, *18*(3), 11-20. https://doi.org/10.1609/aimag.v18i3.1303

Montoya & Peralta v. Character Technologies Inc. [2025] No. 1:25-cv-02907-STV (D. Colo.).

Nilsson, N. J. (2009). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.

Osbeck, L. M., & Held, B. S. (2014). Part II Intuition in Psychology and Cognitive Science. In *Rational Intuition*. Cambridge University Press.

Raine v. OpenAI, Inc. [2025] CGC-25-628528 (Superior Court of California, San Francisco County).

Ratliff, E. (2025, November 12). All of my Employees are AI Agents and so are my Executives. *Wired*. https://www.wired.com/story/all-my-employees-are-ai-agents-so-are-my-executives/

Richards, I. (2025). 'Hypernudging': a Threat to Moral Autonomy?: Author. *Ai and Ethics (Online)*, *5*(2), 1121-1131. https://doi.org/10.1007/s43681-024-00449-y

Romeo, G., & Conti, D. (2026). Exploring Automation Bias in Human–AI Collaboration: a Review and Implications for Explainable AI. AI & Society, 41(1), 259-278. https://doi.org/10.1007/s00146-025-02422-7

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057-1084.

Schmidgall, S., Ziaei, R., Achterberg, J., Kirsch, L., Hajiseyedrazi, S. P., & Eshraghian, J. (2024). Brain-Inspired Learning in Artificial Neural Networks: A Review. *APL Machine Learning*, *2*(2), 021501-021501-021514. https://doi.org/10.1063/5.0186054

Searle, J. R. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, *3*(3), 417-424. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1992). Chapter 4: Consciousness and Its Place in Nature. In *The rediscovery of the Mind*. 83-109. MIT Press.

Theotokis, P. (2025). Human Brain Inspired Artificial Intelligence Neural Networks. *Journal of Integrative Neuroscience*, *24*(4), 26684-26685. https://doi.org/10.31083/JIN26684

Timpe, Kevin. (2008). The Basics. In *Free will : Sourcehood and Its Alternatives* (1st ed., pp. 1-17). Continuum.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433-460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you Need. *Advances in Neural Information Processing Systems 30 - Proceedings of the 2017 Conference*, *2017-*, 5999-6009.

Vatomsky, S. (2017, September 13). When Societies put Animals on Trial**.** *JSTOR Daily*. https://daily.jstor.org/when-societies-put-animals-on-trial/

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Po-Sen, H., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Iason Gabriel. (2021). Ethical and Social Risks of Harm From Language Models. *arXiv.Org*. https://doi.org/10.48550/arXiv.2112.04359

Weizenbaum, J. (1966). ELIZA-a Computer Program for the Study of Natural Language Communication Between man and Machine. *Communications of the ACM*, *9*(1), 36-45. https://doi.org/10.1145/365153.365168